UNIVERSITY OF CALIFORNIA
SANTA CRUZ

FLEXIBLE INTEGRO-DIFFERENTIAL EQUATIONS FOR BAYESIAN MODELING OF SPATIO-TEMPORAL DATA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS AND APPLIED MATHEMATICS

by

Robert Richardson

June 2015

The Dissertation of Robert Richardson
is approved:

_____
Professor Bruno Sansó, Advisor, Chair

_____
Professor Athanasios Kottas, Advisor

_____
Professor Rajarshi Guhaniyogi

_____
Professor Steve Munch

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Flexible Integro-Differential Equations for Bayesian Modeling of Spatio-Temporal Data

by

Robert Richardson

Integro-Differential Equations (IDEs) are a novel way of dynamically modeling spatio-temporal data. IDEs are characterized by a kernel which controls the spatial and temporal associations. The ubiquitous choice for kernel has been Gaussian. We explore advantages of more flexible kernel choices. One-dimensional space is considered initially, replacing the Gaussian IDE kernel with more flexible parametric families of distributions. The kernels are chosen based on stochastic partial differential equation approximations which connect characteristics of the kernel with interpretable physical properties of the underlying process controlling the data. Next, Dirichlet process mixtures of normal distributions are used to model non-parametrically the IDE kernel. Computational issues arise using non-parametric kernels which are solved using Hermite polynomials and Hamiltonian Monte Carlo sampling. To develop flexible modeling in two-dimensional space, we propose bivariate stable distributions as IDE kernels. By using Bernstein polynomials as a prior for the measure defining the bivariate stable, a wide variety of shapes can be achieved. Bivariate stable kernels will be shown to outperform the Gaussian kernel by comparing K-step ahead predictions for Pacific sea surface temperature anomalies. Through study of properties for the proposed models, and empirical investigation with synthetic and real data, we demonstrate that the method-

ology has the potential to significantly improve the inference and forecasting capacity of IDE models based on Gaussian kernels.

Acknowledgments

My family has been amazingly supportive. My wife has helped me stay focused and my kids have woken me up at all hours of the night so I can finish my assignments. My friends Tony and Juan have kept an eye on me, making sure I remember deadlines, and Tracie Tucker has patiently worked with me when I still missed them. My advisers, Brunó and Thanasis, have given me amazing opportunities to jump-start a professional career as well as survived my writing skills.

Chapter 1

Introduction

A spatio-temporal data set refers to data collected across a spatial field and over several time points. Climatological and environmental variables provide several common and abundant examples of data recorded in space and time. In addition to traditional examples of environmental space-time variables, such as temperature or precipitation, there is an increasing ability to store and monitor the dynamics of different types of georeferenced processes. Data for housing costs, crime rates, population growth, soil content, and disease incidence, are some of the many examples of variables that are of interest in areas as diverse as spatial econometrics, epidemiology, and geography, to mention a few.

The field of time series has produced a rich body of literature during at least the last 50 years (Hamilton, 1994; Shumway and Stoffer, 2011). Spatial statistics, despite the seminal work by Matheron (1963), was a fringe area as recently as the early 1990s (Cressie, 1993), but has since received a great deal of attention within the statistical community. Spatio-temporal models stem naturally from these areas, but a systematic treat-

ment of spatio-temporal statistical models has only recently been developed (Cressie and Wikle, 2011). Compared to times series and spatial statistics, the fundamental challenge of spatio-temporal models is to capture the interactions between the spatial and temporal components.

Three general methods are currently used to analyze data from spatio-temporal processes of the form $\{X_t(s) : s \in \mathcal{S}, t \in \mathcal{T}\}$, where s indexes the spatial domain $\mathcal{S}$ and t indexes the time domain $\mathcal{T}$. The first involves an extension of the traditional approach to modeling random fields, that focuses on the first and second moment of the process. The goal is to find general families of space-time correlation functions of the form $Cov(X_t(s), X_u(v)) = C(s, v, t, u)$, which are "smooth everywhere" and yet "allow different degrees of smoothness" (Stein, 2005b). In this setting, both s and t are considered as continuous indexes. This lends flexibility to the models, but requires dealing with potentially large covariance matrices. This approach can thus have important computational drawbacks when large spatial domains or long time periods are considered. Much of the current work in this area is dedicated to developing non-stationary and non-separable covariance structures (Gneiting, 2002; Schmidt and O'Hagan, 2003). A spatio-temporal covariance function is separable if $C(s, v, t, u) = C_1(s, v)C_2(t, u)$ where C_1 is a spatial covariance and C_2 is a temporal covariance. A stationary spatio-temporal covariance function has the property that the spatial and temporal components enter only through the difference between two locations and times, $C(s, v, t, u) = C(s - v, t - u)$. Variations include where only the spatial or only the temporal dependence is stationary. While the computational efficiency of separable and stationary covariance functions has made them useful, they are simply not

realistic for most naturally occurring physical processes. Two additional considerations for spatio-temporal processes are linearity and Gaussianity. Linearity has come to refer to the relationship between $X_t(s)$ and $X_{t-1}(s)$. Non-linear models have been constructed by using interaction terms between different locations at the previous time point to determine the value of the process at the current time point (Wikle and Hooten, 2010). Assuming Gaussian data may be too restrictive for many applications. Gelfand et al. (2005) and Kottas et al. (2008) use fully flexible error distributions to model spatial and spatio-temporal data. These extensions are difficult to achieve with covariance modeling alone.

A second common modeling approach for spatio-temporal data is an extension of deterministic dynamical models that incorporates stochastic components. This leads to stochastic partial differential equation (SPDE) models. For instance, Jones and Zhang (1997) consider the SPDE $\frac{\partial}{\partial t}X_t(s) - \beta\frac{\partial^2}{\partial s^2}X_t(s) + \alpha X_t(s) = \delta_t(s)$, where $\delta_t(s)$ is a zero mean error process. This SPDE is called a diffusion-injection equation and is just one of the various SPDE-based models commonly used for naturally occurring physical processes. The deterministic relationships which motivate SPDEs will often involve non-linear components (Hooten and Wikle, 2008).

The third method is to obtain an explicit description of the dynamics of the process by specifying its evolution as a function of the spatial distribution of the process. A dynamic spatio-temporal model can be written as

$$X_t(s) = \mathcal{M}(X_{t-1}(s), s, \boldsymbol{\theta}) + \omega_t(s), \quad t = 1, \ldots, T,$$

where $\mathcal{M}$ represents a specific model configuration, governing the transfer of information from time $t-1$ to time t. Here, $\boldsymbol{\theta}$ is a parameter vector, and $\varepsilon_t(s)$ is a zero mean noise

process which may have a spatially dependent covariance structure. In these models, the process evolves as an entire spatial field over a discrete time component. Cressie and Wikle (2011) strongly support this approach, and suggest a "hierarchical dynamical spatio-temporal model" of the form

$$\boldsymbol{Y}_t = \boldsymbol{H}_t\boldsymbol{X}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{V}_t), \quad t = 1, \ldots, T \tag{1.1}$$

$$\boldsymbol{X_t} = \mathcal{M}_t(\boldsymbol{X}_{t-1}, \boldsymbol{\theta}) + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\boldsymbol{0}, \boldsymbol{W}_t), \quad t = 1, \ldots, T, \tag{1.2}$$

where $\boldsymbol{Y}_t$ is the vector of data, $\boldsymbol{X}_t$ a vector of latent variables representing an underlying process that is linked to $\boldsymbol{Y}_t$ through the incidence matrix $\boldsymbol{B}_t$. Moreover, $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\omega}_t$ are noise terms with specified covariances $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$, respectively.

A specific case of the model described by equations (1.1) and (1.2) is the integro-difference equation (IDE) spatio-temporal model. We consider IDE models of the form

$$X_t(s) = e^{\lambda} \int k(u|s, \boldsymbol{\theta}) X_{t-1}(u) du + \omega_t(s), \tag{1.3}$$

where $k(\cdot)$ is a redistribution kernel with parameter vector $\boldsymbol{\theta}$, and $\omega_t(s)$ is an error process which may be spatially colored. This kernel weights the contribution of the process at time $t-1$ to the process at time t at location s. The scaling term λ controls the growth or decay of the process. Typically, the center of the kernel for each location is somewhere near s, resulting in nearby values being weighted more heavily than others. The spatial dependency in the IDE model arises from nearby observations sharing large contributions from many of the same observations of the previous time point. Thus, the spatial and temporal relationships interact with each other as the process evolves.

Originally used by ecologists studying the growth and spread of species (Kot et al., 1996), integro-difference equations were introduced for general spatio-temporal processes in

Wikle and Cressie (1999). In Wikle (2002) the IDE kernel is specified parametrically through a Gaussian distribution with unknown location and scale parameters. The stochastic properties of the process that results from an IDE, such as stationarity and separability, are explored in Brown et al. (2000) and Storvik et al. (2002). An important extension where the parameters of the kernel are spatially indexed is presented in Wikle (2002) and Xu et al. (2005).

Overall, the literature is dominated by IDE models based on Gaussian kernels. Though there is some mention of non-Gaussian kernels, it is without exploring the modeling benefits and inferential issues arising from the use of wider kernel families. Spatio-temporal data can have a variety of features that may not be represented well by a Gaussian kernel IDE model. In this report, we focus on the exploration of the properties and the development of inferential methods to deal with non-Gaussian kernel IDE models. We will show that, for hierarchical models as in equations (1.1) and (1.2), an IDE with a kernel more flexible than the Gaussian can lead to improved model performance and prediction, and capture a wider array of process dynamics. Initially, this extension will be limited to one-dimensional kernels to more thoroughly compare the models with different kernels and to match some theoretical results which are presented.

Chapter 2 includes a review of IDE modeling. In this chapter, details of fitting spatio-temporal data with IDE models are specified. Chapter 3 introduces two parametric alternatives to the Gaussian kernel: the asymmetric Laplace and the stable family. These kernels allow a process following an IDE structure to maintain certain physical characteristics. We show that prediction and accuracy can be improved by using a more flexible

kernel. Chapter 4 builds upon the idea that more flexibility yields better models and places a non-parametric prior on the kernel. Computational solutions are presented which account for the difficulty in learning the high-dimensional kernel parameter set in this setting. When the kernel can be learned properly, IDE models with spatially varying non-parametric kernels are very promising. The prospect of learning non-parametric kernel parameters when extending to 2-dimensional space seems to be too ambitious. The solution is to achieve a wide variety of kernel shapes by semi-parametrically modeling the measure which controls the shape of a bivariate stable distribution. Details of fitting IDE models with bivariate stable kernels are shown in Chapter 5.

Chapter 2

Review of IDE Modeling

Using integro-difference equations to model spatio-temporal data is relatively new. Initial approaches did not require the kernel to be a density, but rather decomposed the kernel into a basis function expansion and estimated the coefficients directly (Wikle and Cressie, 1999). While this non-parametric approach to modeling the kernel may have produced accurate results, it reveals no information about the physical process controlling the data. In Brown et al. (2000) and Storvik et al. (2002), a process following a Gaussian kernel IDE model is studied carefully. It was shown that the process has an equivalent stochastic partial differential equation representation and has a stationary covariance function. In this chapter we review these results and extend them to wider classes of kernels. We also review the general methodology associated with fitting IDE models and learning the kernel parameters. In chapters 3-5, the kernel is modeled in different ways, but the actual method of fitting the data will remain mostly the same. The methodology described in this chapter will be referred to in future chapters and can be used for general IDE modeling of

spatio-temporal data.

2.1 Stationarity of IDE Models

Due to the linearity of IDEs, the expected value and covariance of a process following an IDE without noise are $E\left[X_t(s)\right] = \int k(u|s,\boldsymbol{\theta}_s)E[X_{t-1}(u)]du$ and $Cov(X_t(s), X_t(r)) = \int k(u|s,\boldsymbol{\theta}_s)k(v|r,\boldsymbol{\theta}_r)Cov(X_{t-1}(u), X_{t-1}(v))dudv$. If the IDE process at time $t-1$ and location u has mean function $m_{t-1}(u)$ and covariance function $\rho_{t-1}(u,v)$, then the mean process for time t is $m_t(s) = \int k(u|s,\boldsymbol{\theta}_s)m_{t-1}(u)du$ and the covariance function is $\rho_t(s,r) = \int\int k(u|s,\theta_s)k(v|r,\boldsymbol{\theta}_r)\rho_{t-1}(u,v)dudv$. Brown et al. (2000) showed that IDE models are stationary in space, but the work was limited to kernels with parameters which were not spatially varying. That work is reviewed here in the context of spatially varying kernels.

Lemma 1. *Consider an IDE process, $X_t(s)$, with a stationary initial process $X_0(s)$. Then, consider a family of kernels belonging to a location family of distributions. The process will be stationary for all $t > 0$ only when the parameters of the kernel do not depend on the location s.*

Proof. Assume $Cov[X_{t-1}(s), X_{t-1}(r)] = \rho(|s-r|)$, a stationary covariance function and that the error process $\omega_t(s)$ is also spatially stationary with covariance function $\gamma(|s-r|)$. Then,

$$\begin{aligned} Cov[X_t(s), X_t(s+r)] &= Cov\left[\int k(u|s,\boldsymbol{\theta}_s)X_{t-1}(u)du + \omega_t(s), \int k(v|s+r,\boldsymbol{\theta}_{s+r})X_{t-1}(v)dv + \omega_t(s+r)\right] \\ &= \int\int k(u|s,\boldsymbol{\theta}_s)k(v|s+r,\boldsymbol{\theta}_{s+r})\rho(|u-v|)dudv + \gamma(r). \end{aligned}$$

Using the transformations $\eta = u - v$ and $w = v - s$, when $k(u|s,\boldsymbol{\theta}_s) = k(u-s|\boldsymbol{\theta}_s)$, the

covariance function simplifies to

$$Cov[X_t(s), X_t(s+r)] = \int k(\eta + w|\boldsymbol{\theta}_s)k(w - r|\boldsymbol{\theta}_{s+r})\rho(|\eta|)d\eta dw + \gamma(r)$$

The covariance is a function of s and r, which implies non-stationarity. However, when the parameter vector does not depend on the location, meaning $\boldsymbol{\theta}_s = \boldsymbol{\theta}$, the location s disappears from the covariance and the process is stationary. □

This result relies on the kernel belonging to a location family, which is the case for all modeling examples that we are aware of. According to the lemma, even very complicated kernels will yield stationary processes when the parameters are constant for every location.

The temporal stationarity of an IDE model is explored in both Brown et al. (2000) and Storvik et al. (2002), and we review the latter's explanation. In order for the process to have spatio-temporal stationarity, the spatial covariance at each time point must be the same. Even when the spatial process at each time point is stationary, the covariance structure may change from one time point to the next. If the kernels are location family kernels and the process is spatially stationary at each time point then the covariance function at time t between the process at locations s and $s+r$ is $\rho_t(r) = \int\int k(u|\theta_s)k(v|\boldsymbol{\theta}_{s+r})\rho_{t-1}(r+v-u)dudv + \gamma(r)$. Let $f *_r g$ denote a convolution about r of the functions f and g. Then the covariance can be rewritten as $\rho_t(r) = k *_r k *_{-r} \rho_{t-1}(r) + \gamma(r)$. Let $H(\omega), \Gamma(\omega)$, and $R(\omega)$ be the inverse fourier transform of the kernel, $k(u)$, the error covariance, $\gamma(r)$, and covariance of the process, $\rho_t(r)$, respectively. When $\rho_t(r) = \rho(r)$ for all t, the inverse Fourier transform of the covariance of the process must be

$$R(\omega) = \frac{\Gamma(\omega)}{1 - H(\omega)H(-\omega)} \tag{2.1}$$

This is the spectral density of the covariance of the process, and it corresponds to a unique covariance function, which will be the Fourier transform of $R(\omega)$. The implication is that, while stationarity can be achieved for a specific covariance function of the process, IDE modeling is not restricted to stationarity. Even though the stationary covariance is a function of the kernel, the ability to model non-stationarity in time is not kernel-specific. By choosing the kernel and error process carefully, the covariance structure of the process can be reconstructed by the Fourier transform of $R(\omega)$ in equation (2.1), if it is stationary. If the inverse Fourier transforms of the kernel, error covariance, and process covariance yield an inequality in (2.1), the process will be non-stationary.

2.2 PDE Approximations

Partial differential equations have provided a powerful way to represent scientifically motivated relationships. This section reviews and expands on the connection between PDEs and IDEs. Two different PDE representations of the IDE process will be shown.

2.2.1 High Order Moments SPDE Representation

Brown et al. (2000) consider an IDE model where the time increment is infinitesimal. This is only possible when the kernel is infinitely divisible, which means that for any integer n, there exist n identically distributed random variables whose sum is a random variable belonging to the kernel family of distributions. Using Taylor series expansions, the solution of the IDE in equation (1.3), when the kernel is infinitely divisible and from a

location family, satisfies the approximation

$$\frac{\partial X_t(s)}{\partial t} \approx \lambda X_t(s) - \mu \frac{\partial X_t(s)}{\partial s} + \frac{1}{2}\sigma^2 \frac{\partial^2 X_t(s)}{\partial s^2} + B_t(s) \tag{2.2}$$

where μ and σ^2 are, respectively, the mean and the variance of the kernel, and $B_t(s)$ is Brownian motion. A distribution, $\mathcal{F}$, is infinitely divisible if any random variable $X \sim \mathcal{F}$ can be written as $X = \sum_{i=1}^{n} X_i$, for any n, where X_i are identically distributed random variables (Steutel and Harn, 2003). Intuitively, the effect of an infinitely divisible kernel controlling the evolution of an IDE for one unit of time can be decomposed into the sum of the effects of n IDEs operating on $1/n$ units of time. Thus, infinite divisibility allows a discrete time IDE to be approximated by an SPDE, which is a continuous time model. The model in equation (2.2) depends on two parameters, μ and σ^2, that control, respectively, the advection and diffusion of the process $X_t(s)$. Thus, the SPDE approximation of the IDE sheds light on how the kernel parameters control the physical properties of the process $X_t(s)$.

Following the framework in Brown et al. (2000), we establish the following result that provides an SPDE representation of an IDE using moments of order higher than two.

Lemma 2. *Consider the IDE model in* (1.3) *with* $\lambda = 0$*, and with an infinitely divisible kernel from a location family for which the first* J *central moments,* $\mu_1, \ldots, \mu_J$*, exist. Moreover, assume that* $\frac{\partial^j}{\partial s^j} X_{t-\delta}(s)$ *exists for any (small)* $\delta > 0$ *and for* $j = 1, ...J$*. Then, the solution to the IDE equation can be approximated by the solution of the equation*

$$\frac{\partial X_t(s)}{\partial t} \approx \sum_{j=1}^{J} (-1)^j \frac{1}{j!} \mu_j \frac{\partial^j X_t(s)}{\partial s^j} + B_t(s) \tag{2.3}$$

where $B_t(s)$ *is Brownian motion.*

Proof. For an infinitely divisible location kernel $k(s-u|\boldsymbol{\theta})$, we define $k_{\frac{1}{n}}(s-u|\boldsymbol{\theta}_{\frac{1}{n}})$ as an n-fold self convolution, $k_{\frac{1}{n}}(x) * k_{\frac{1}{n}}(x) * ... * k_{\frac{1}{n}}(x) = k(x)$, and $\boldsymbol{\theta}_{\frac{1}{n}}$ as the adjusted parameter set induced by the self-convolution. Let $\delta = 1/n$. Then, by representing the process at time $t-\delta$ as a Taylor series with J terms, we can write $X_t(s)$ as

$$\begin{aligned}\int k_\delta(u|\boldsymbol{\theta}_\delta) X_{t-\delta}(s-u)du + \omega_{t,\delta}(s) &= \int k_\delta(u|\boldsymbol{\theta}_\delta)\left[\sum_{j=0}^{J}(-1)^j \frac{1}{j!}u^j \frac{\partial^j}{\partial s^j}X_{t-\delta}(s) + o(u^J)\right]du + \omega_{t,\delta}(s) \\ &\approx X_{t-\delta}(s) + \sum_{j=1}^{J}(-1)^j \frac{1}{j!}E_{k_\delta}\left[u^j\right]\frac{\partial^j}{\partial s^j}X_{t-\delta}(s) + \omega_{t,\delta}(s)\end{aligned}$$

where E_{k_δ} is the expected value with respect to the distribution with density k_δ, and $\omega_{t,\delta}(s)$ is the transformed error process. By rearranging terms and dividing by δ we have

$$\frac{X_t(s) - X_{t-\delta}(s)}{\delta} = \sum_{j=1}^{J}(-1)^j \frac{1}{j!}\left(\mu_j + h(\delta)\right)\frac{\partial^j}{\partial s^j}X_{t-\delta}(s) + \frac{1}{\delta}\omega_{t,\delta}(s).$$

where $h(\delta)$ is a polynomial in δ with lowest order of 1. As $\delta \to 0$, the function $h(\delta) \to 0$ and we are left with the desired result. □

The first two moments of the kernel control the advection and diffusion of the resulting process. Extra-diffusive dynamics depend on even terms of order higher than two. The third moment is known to control dispersion, which in this context allows for extra variability in how the process behaves from one spatial location to the next. Lemma 2 suggests that a more flexible kernel can model more complicated dynamics.

2.2.2 Hazard Function PDE Representation

An alternative characterization of an IDE in terms of a differential equation has been studied by ecologists dealing with the dispersal of organisms after their introduction in a foreign region (Neubert et al., 1995). A simple experiment would consist of a researcher placing a foreign species in the middle of an open field. After a specified period of time,

they would return and measure how the plant spread over the field. They would use what is essentially an one time step IDE process to describe the behavior of how the organism spread.

Lemma 3. *Let k, S, and h be, respectively, the kernel density, and the corresponding survival and hazard functions defined as $S(s) = 1 - \int_{-\infty}^{s} k(u)du$ and $h(s) = k(s)/S(s)$. Then, setting the initial condition to $u_0(s) = X_{t-1}(s)$, the system of differential equations*

$$\frac{\partial u_\tau(s)}{\partial \tau} = -\frac{\partial u_\tau(s)}{\partial s} - h(\tau)u_\tau(s) \quad \text{and} \quad \frac{\partial v_\tau(s)}{\partial \tau} = h(\tau)S(0)u_\tau(s) \tag{2.4}$$

has the solution

$$u_\tau(s) = X_{t-1}(s-\tau)\frac{S(\tau)}{S(0)} \quad \text{and} \quad v_\tau(s) = \int X_{t-1}(s-u)k(u)du\ .$$

Proof. We confirm that this is an IDE representation using the method of characteristics. To use this method, we find curves where the PDE is trivial and then create functions of those curves based on the initial conditions. The characteristics curves can be found by solving the differential equations $d\tau = ds$ and $d\tau = -(h(\tau)u)^{-1}du$. The first PDE is simple to integrate both sides. The second can be solved for $u = C\exp\left[-\int h(\tau)d\tau\right]$.

According to the method of characteristics, the general solution can be written as $u = g(s-\tau)\exp\left[-\int h(\tau)dt\right]$. Neubert et al. (1995) assume the initial condition $u(s,0) = \delta(s)$ because all the organisms begin in one location, but in general we can use the initial condition $u(s,0) = X_t(s)$, that is, our initial condition is the process at the previous time. Using properties of hazard functions, based on this initial condition, the general function for $u(s,\tau)$ is $X_{t-1}(s-\tau)S(\tau)/S(0)$ where $S(\cdot)$ is the survival function. Solving for v proceeds

by integrating both sides of $\frac{dv}{d\tau} = X_{t-1}(s-\tau)\frac{S(\tau)}{S(0)}h(\tau)S(0)$, which then becomes

$$v(s,\tau) = \int X_{t-1}(s-\tau)k(\tau)d\tau.$$

If the initial condition for $u(x,0)$ is $X_{t-1}(s)$, then the solution for $v(s,\tau)$ is $X_t(s)$. □

In ecology, the interpretation of $u_\tau(s)$ is that of a latent process representing the path of particulates in motion. The process $v_\tau(s)$ is a measure of the organisms once they have settled. The variable τ is an index of the path of the process in-between time steps. As τ travels from 0 to ∞, the process $X_t(s)$ moves from time t to time $t+1$, and the process $u_\tau(s)$ becomes 0 as all the particulates settle into locations contributing to $v_\tau(s)$. To make this a multi-step process we set the initial value for $u_\tau^{(t)}(s)$ equal to $X_{t-1}(s)$ and solve the series of differential equations $\{(u_\tau^{(t)}(s), v_\tau^{(t)}(s)) : t = 1, ..., T\}$ piece by piece.

From the above discussion, we can identify $v_\tau(s)$ with $X_t(s)$, implying that the dynamics of a process that satisfies an IDE with no random shocks, are regulated by the PDE in equation (2.4). This indicates that the behavior of an IDE process depends on the hazard function associated with the kernel. Tail behavior and hazard functions are directly related. This is illustrated in Figure 2.1 for three densities with different tails. Thus, we expect that a kernel with thick tails, such as a Cauchy, will produce solutions to the IDE that behave very differently than those that correspond to a Gaussian kernel IDE.

Lemmas 2 and 3 indicate that there is merit in using IDE kernels with more general high order moments and tail behavior than the Gaussian kernel. In line with the results considered in this section, we seek alternative kernels that belong to infinitely divisible, location families of distributions that possess higher order moments and/or have more flexible tails than the normal. Two parametric families that offer flexibility along these

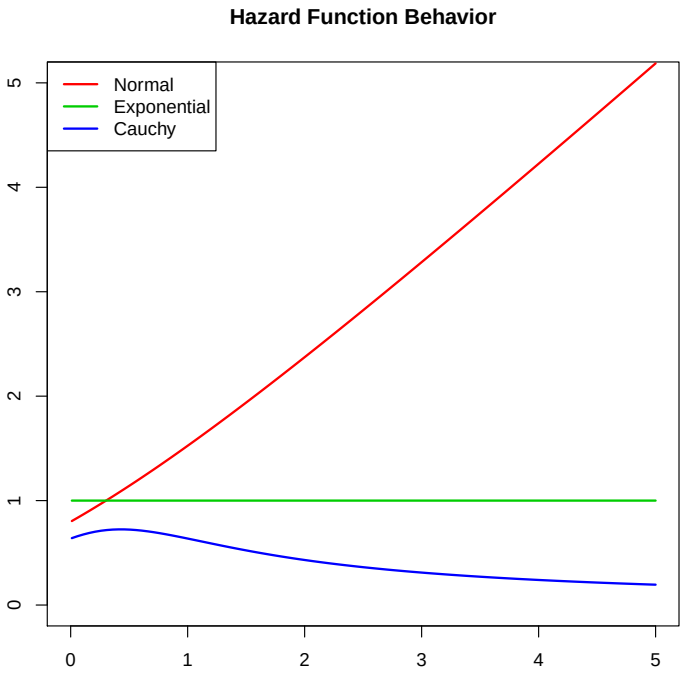


Figure 2.1: The hazard functions for the standard normal, standard Cauchy, and exponential distributions are shown. The Cauchy has polynomial tails that yield a decreasing hazard function. The normal distribution has a hazard function which is increasing, and the exponential hazard is constant.

lines, without compromising tractability, will be presented in Chapter 2.

2.3 Basis Expansion for Model Fitting

We will use an orthogonal basis expansion for both the kernel and the process, where the basis functions, $\{\psi_1, \psi_2, ...\}$, are common to both. In particular,

$$X_t(s) = \sum_{i=1}^{\infty} \psi_i(s) a_i(t) \qquad \text{and} \qquad k(u|s, \boldsymbol{\theta}) = \sum_{j=1}^{\infty} b_j(s, \boldsymbol{\theta}) \psi_j(u), \tag{2.5}$$

where $a_i(t)$ are coefficients for the basis expansion of the process, and $b_j(s, \boldsymbol{\theta})$ are coefficients for the basis expansion of the kernel at location s. Using a set of basis functions where truncation is appropriate, both series in equation (2.5) may be truncated to the first K terms. The value for K should be sufficiently large for the basis expansion to accurately approximate both the kernel and the process.

Due to the orthogonality of the basis functions, the components of the integral in equation (1.3) can be replaced with the basis expansions in equation (2.5) and rewritten as $\int k(s-u|\boldsymbol{\theta})X_t(u)du = \mathbf{a}_t'\mathbf{b}(s,\boldsymbol{\theta})$, where $\mathbf{a}_t = (a_1(t), ..., a_K(t))'$ and $\mathbf{b}(s,\boldsymbol{\theta}) = (b_1(s,\boldsymbol{\theta}), ..., b_K(s,\boldsymbol{\theta}))'$. Moreover, by placing $X_{t+1}(s) = \sum_{i=1}^{K} \psi_i(s)a_i(t+1)$ into the left side of equation (1.3), we obtain $\mathbf{a}_{t+1}'\boldsymbol{\psi}(s) = \mathbf{a}_t'\mathbf{b}(s,\boldsymbol{\theta}) + \omega_{t+1}(s)$, where $\boldsymbol{\psi}(s) = (\psi_1(s), ..., \psi_K(s))'$.

The values for $b_j(s,\boldsymbol{\theta})$ are deterministic, given the choice of kernel and the corresponding parameters. The values for $a_i(t)$ are unknown and vary with time. Under the basis expansion, a data vector, $\boldsymbol{Y}_t$, can be summarized hierarchically as follows:

$$\boldsymbol{Y}_t = \boldsymbol{\Psi}\mathbf{a}_t + \boldsymbol{\varepsilon}_t \tag{2.6}$$

$$\boldsymbol{\Psi}\mathbf{a}_t = \boldsymbol{B}_\theta\mathbf{a}_{t-1} + \boldsymbol{\omega}_t, \tag{2.7}$$

where $\boldsymbol{Y}_t = (Y_t(s_1), ..., .Y_t(s_n))'$ is a noisy realization from the process at time t, $\boldsymbol{B}_\theta = (\mathbf{b}(s_1,\boldsymbol{\theta}) \ldots \mathbf{b}(s_n,\boldsymbol{\theta}))$ is a matrix whose columns consist of the vectors of the kernel basis coefficients, and the (i,j)th element of $\boldsymbol{\Psi}$ is $\psi_i(s_j)$. The vectors $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\omega}_t$ account for observational error and process error, respectively. This representation of an IDE model becomes a state space model by removing $\boldsymbol{\Psi}$ from the left side of equation (2.7). If a discrete orthonormal basis is used, the matrix $\boldsymbol{\Psi}'\boldsymbol{\Psi}$ is equal to the identity matrix, then $\mathbf{a}_t = \boldsymbol{\Psi}'\boldsymbol{B}_\theta\mathbf{a}_{t-1} + \boldsymbol{\Psi}'\boldsymbol{\omega}_t$. For continuous bases, the orthogonality of $\boldsymbol{\Psi}$ is not guaranteed.

Then we create a state space model by rewriting equation (2.7) as $\mathbf{a}_t = (\boldsymbol{\Psi}'\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\boldsymbol{B}_\theta\mathbf{a}_{t-1} + (\boldsymbol{\Psi}'\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\boldsymbol{\omega}_t$.

2.3.1 Selecting the Appropriate Basis

If we are to assign physical interpretations to the estimated kernel, such as the ones implied by (2.3) and (2.4), then the kernel must be appropriately approximated. Choosing the basis used to decompose the IDE model and selecting the number of basis functions must be carefully considered. Depending on the basis chosen, there may be additional considerations. Accuracy improves with a larger number of basis functions, but the size of the latent state vector, $\boldsymbol{a}_t$, in equations (2.6) and (2.7) increases with the number of basis functions. The result is an inverse relationship between accuracy and computational speed.

Fourier Basis

For a bounded spatial domain, say $[r_1, r_2]$, it is natural to consider the orthogonal family given by the Fourier basis. In such case, the kernel need only be specified through its characteristic function. For example, a Gaussian kernel has Fourier coefficients $b_{2j-1}(s, \boldsymbol{\theta}) = r^{-1/2}\exp(-.5\rho_j^2\sigma^2)\cos(\rho_j(s+\mu))$, and $b_{2j}(s, \boldsymbol{\theta}) = r^{-1/2}\exp(-.5\rho_j^2\sigma^2)\sin(\rho_j(s+\mu))$, where $r = r_2 - r_1$ and $\rho_j = 2\pi j/r$ is the spatial frequency. Both the number of basis functions and the spatial domain must be specified. One major issue is that if the kernel parameters change, the optimal number of basis functions changes as well. Figure 2.2 shows that kernels with a smaller variance are more difficult to approximate well with a small number of basis functions. Calling the variance small or large is relative to the range $[r_1, r_2]$. For example, in Figure 2.2 the same normal density is approximated in the left and right images, but the

range has changed. If some information about the kernel width in relation to the range of the data is known, it can be used to estimate the optimal number of basis functions to be used. However, if it is unknown, it may be safer to select a larger number of basis functions and check the posterior to see how appropriate the choice was. Another factor to consider is

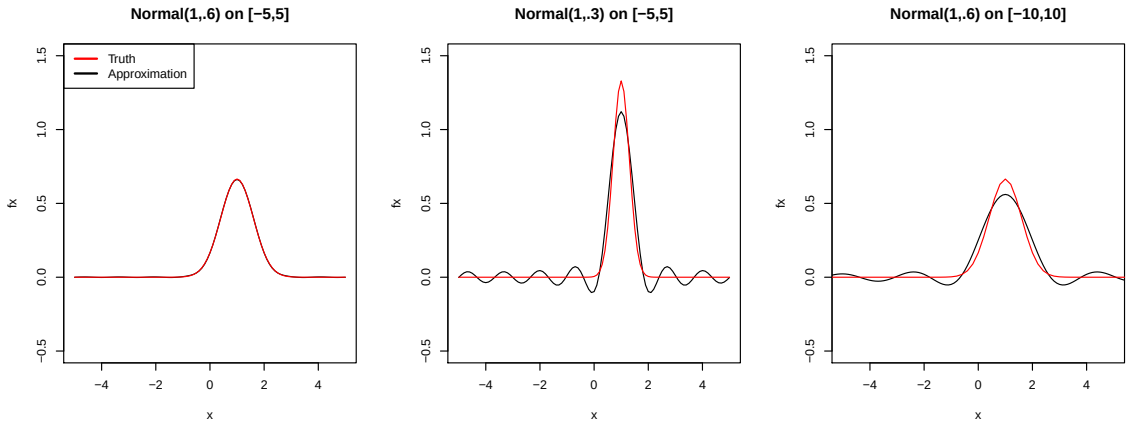


Figure 2.2: 15 Basis functions are used to approximate a normal density. On the left, the Fourier approximation works well for a normal density with a larger variance and a Fourier period of 10. When the variance is smaller (middle), or when the period is larger (right), the approximation is much worse.

the periodicity implied by a Fourier basis. If the density function contains significant mass outside the defined region it will wrap around the other edge of the region. This can be seen in Figure 2.3 where the range is -2 to 2 and the density extends past the range. When this happens, the IDE integral in equation (1.3) will artificially give weight locations which are on the opposite end of the spatial domain. To avoid this, the values for the range r_1 to r_2 should be chosen to be larger than the actual bounds of the data. For example, if the data locations extend from -4 to 4, the period of the Fourier approximation should be from

-5 to 5, or when a larger kernel is used, from -6 to 6. To make many of these calculations more simple, the data locations can be centered around 0 to make $r_2 = -r_1$.

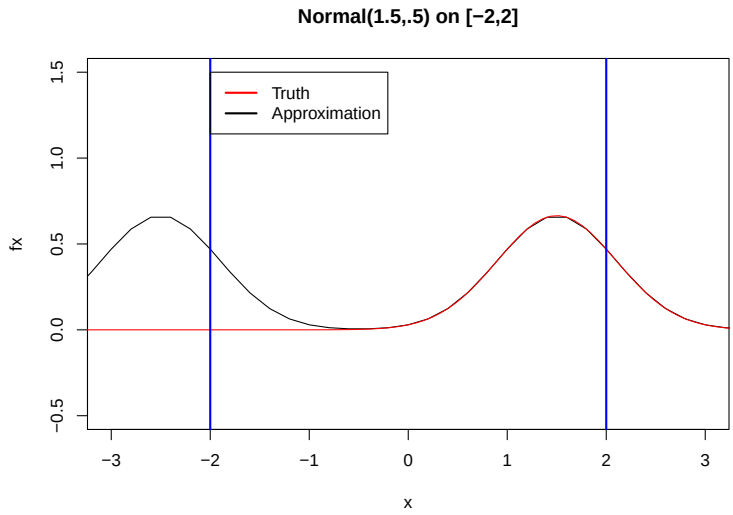


Figure 2.3: A normal density with mean 1.5 and standard deviation .5 is approximated with Fourier basis coefficients on the range of -2 to 2. The effect of the implied periodicity is shown.

Empirical Orthogonal Functions

A method of reducing the dimensionality of spatial processes is to use Empirical Orthogonal Functions (EOF). This involves a reduction of the data based on the principal components. The continuous analog of this, known as the Karhunen-Loeve expansion can be shown to minimize squared mean distance between the target process and the approximation. In practice, one would determine an empirical covariance matrix from data lying on a regular grid and finding the eigenvalues $\lambda_1, ..., \lambda_N$ and eigenvectors $e_1, ..., e_N$. The eigenvectors form a basis which can represent the data. The amount of variation in N

observations explained by n eigenvectors where $n < N$ is $100\% \times \sum_{i=1}^{n} \lambda_i / \sum_{j=1}^{N} \lambda_j$. To approximate a kernel with an EOF basis, one would need to discretize the kernel density function to the same grid as the data and apply a discrete transform. Choosing the number of basis functions may be simpler in this case, because the number of basis functions to be chosen may relate to a desired percent of variation explained in the data. Because this is a data-based basis, the uncertainty introduced by the approximation will typically not propagate.

Other Basis Function Choices

There are a number of other basis function choices which could be made, such as wavelets or splines. Orthogonal polynomial basis functions could also be an effective basis for modeling the kernel accurately in a reasonable number of basis functions. A polynomial basis $P_1, P_2, ...$ is orthogonal with respect to a weight function $w(x)$ when the inner product, $< P_i, P_j >_w = \int P_i(x)P_j(x)w(x)dx$, is 0 when $i \neq j$ and is not 0 when $i = j$. There are a variety of these for a number of weight functions including Legendre polynomials for $w(x) = \mathbb{1}_{[0,1]}(x)$, Laguerre polynomials when $w(x) = e^{-x}\mathbb{1}_{[0,\infty)}(x)$, and Hermite polynomials when $w(x) = e^{-x^2}$. From the equations in (2.4), it can be seen that tail behavior may have a significant impact on the process. Approximating this tail behavior accurately may be more difficult for standard basis choices, but perhaps using an orthogonal polynomial basis where the weight function mirrors the tail behavior will improve accuracy in approximating these kernel densities. For example, the Gaussian kernel has squared exponential tails so using Hermite polynomials where the weight function is a squared exponential may be a very natural way to represent the density. Likewise with

an exponential kernel and the Laguerre polynomials. Many of the basis function choices mentioned here do not require a range to be specified as in the Fourier basis, but do require the number of basis functions to be chosen carefully. Again, the number of basis functions required will depend on kernel width.

2.4 MCMC Details

This section details the procedure used for learning model parameters in an IDE spatio-temporal model. As outlined in Cressie and Wikle (2011) and summarized in equations (1.1) and (1.2), we use a hierarchical dynamic linear model framework. For a data vector, $\boldsymbol{Y}_t = (Y_t(s_{t,1}), ..., Y_t(s_{t,n_t}))'$, using a basis function expansion, the full IDE model can be written as

$$\begin{aligned} \boldsymbol{Y}_t|\boldsymbol{a}_t, \sigma^2 &\sim N(\boldsymbol{\Psi}_t\boldsymbol{a}_t, \sigma^2\boldsymbol{I}_{n_t}), \quad t = 1, ..., T \qquad (2.8) \\ \boldsymbol{a}_t|\boldsymbol{a}_{t-1}, \tau^2, \boldsymbol{\theta} &\sim N(\boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{a}_{t-1}, \boldsymbol{G}_t\boldsymbol{W}_t\boldsymbol{G}_t') \qquad (2.9) \\ \boldsymbol{\theta}|\gamma &\sim p(\boldsymbol{\theta}|\gamma), \quad \sigma^2, W_t \sim p(\sigma^2)p(W_t), \qquad (2.10) \end{aligned}$$

where $\boldsymbol{a}_t$ are the latent state variables representing the stochastic basis coefficients of the process and $\boldsymbol{G}_t = (\boldsymbol{\Psi}_t'\boldsymbol{\Psi}_t)^{-1}\boldsymbol{\Psi}_t'$. The length of the state vector is equal to the number of basis functions. The observational variance is $\sigma^2\boldsymbol{I}_{n_t}$ and the variance of the process level is $\boldsymbol{G}_t\boldsymbol{W}_t\boldsymbol{G}_t$. The (i,j)th element of $\boldsymbol{\Psi}_t$ is $\psi_i(s_{t,j})$, the i-th basis function evaluated at $s_{t,j}$ and the (i,j)-th element of $\boldsymbol{B}_{\theta,t}$ is $b_i(s_{t,j}|\boldsymbol{\theta})$, the i-th basis coefficient of the kernel at the location $s_{t,j}$. In full generality, the dimension of these matrices are time-indexed because the locations at which the data occurs may change. The parameters involved in the model are σ^2, $\boldsymbol{W}_t$, $\boldsymbol{\theta}$, and possibly γ if hyper-priors are used. The prior for the kernel parameter

set $\boldsymbol{\theta}$ depends on the family of distributions chosen for the kernel. For example, a Gaussian kernel parameter vector includes the mean and variance of the distribution, and a normal prior could be placed on the mean and a Gamma or inverse Gamma prior could be placed on the variance. The posterior for σ^2 will be conjugate if the prior is inverse gamma with parameters α_σ and β_σ. We also need to define a prior for $\boldsymbol{a}_0$, the coefficients of the basis expansion of the time zero process. This will be a multivariate normal with mean $\boldsymbol{m}_0$ and variance $\boldsymbol{C}_0$. The priors for $\boldsymbol{a}_t$ for $t > 0$ are stated in equation (2.9).

The spatial locations used to determine $\boldsymbol{\Psi_t}$ in the observation equation shown in equation (2.8) must be the observed locations of the data. The number of locations and where they occur do not need to be the same for each time point. For simplification in computation, we can select a grid to obtain a representation for the evolution equation which is different than that in the observation equation. A new grid can be defined on points $r_1, ..., r_n$, which is the same for all time points. Then set $\boldsymbol{G}_t = \boldsymbol{G} = \boldsymbol{(\Psi'\Psi)^{-1}\Psi'}$, where $\boldsymbol{\Psi}$ is determined using the new grid. Also, $\boldsymbol{B}_{\theta,t} = \boldsymbol{B}_\theta$ is determined using the new grid. The advantage of using locations which are constant in time for the process level is that populating the matrix $\boldsymbol{B}_{\theta,t}$ can be time consuming for a complicated kernel distribution. Over the course of a lengthy Monte Carlo algorithm, the speed-up of simplifying to a single $\boldsymbol{B}_\theta$ may be significant.

Utilizing Gibb's sampling, this model becomes a conditional dynamic linear model. When $\boldsymbol{\theta}, \sigma^2$, and $\boldsymbol{W}_t$ are known, methods to sample from the posterior of the state variables have been developed in the literature. Specifically, we will use Forward Filtering Backwards Sampling (West and Harrison, 1997) to draw a sample for $\boldsymbol{a}_0, ..., \boldsymbol{a}_T$. Conditional on sampled

state vectors, the posterior for σ^2 is

$$\sigma^2|\cdot \sim IG\left(\alpha_\sigma + \frac{nT}{2}, \beta_\sigma + \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{\Psi}_t\boldsymbol{a}_t)'(\boldsymbol{Y}_t - \boldsymbol{\Psi}_t\boldsymbol{a}_t)\right) \tag{2.11}$$

The set of kernel parameters, $\boldsymbol{\theta}$, must be estimated using a Metropolis-Hastings algorithm or something similar. The conditional posterior of $\boldsymbol{\theta}$ is proportional to $\prod_{t=1}^{T} p(\boldsymbol{a}_t|\boldsymbol{a}_{t-1}, \boldsymbol{W}_t, \boldsymbol{\theta})p(\boldsymbol{\theta}|\gamma)$. For applications with a complicated kernel, such as a Dirichlet process mixture of normals, learning the parameters may be very difficult using standard Metropolis-Hastings. For situations such as these, a Hamiltonian Monte Carlo algorithm (Neal, 2011) may be used to sample from the posterior distribution of $\boldsymbol{\theta}$.

Estimating the states is done using standard filtering formulas as found in Prado and West (2010). First the prior mean and covariance estimates for $\boldsymbol{X}_0$ must be set as $\boldsymbol{m}_0$ and $\boldsymbol{C}_0$. Let $\boldsymbol{Q}_t = \boldsymbol{G}_t\boldsymbol{W}_t\boldsymbol{G}_t$ and $\boldsymbol{R}_t = \sigma^2\boldsymbol{I}_{n_t}$ Given all information up to time t, denoted as $\boldsymbol{D}_t$, the posterior distributions of the state vectors $\boldsymbol{X}_t|\boldsymbol{D}_t$ are $N(\boldsymbol{m}_t, \boldsymbol{C}_t)$. These can be found using recursive formulas

$$\begin{aligned}
\boldsymbol{m}_t &= \boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{m}_{t-1} + \boldsymbol{K}_t\left(\boldsymbol{Y}_t - \boldsymbol{\Psi}_t\boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{m}_{t-1}\right) \\
\boldsymbol{C}_t &= \left(I - \boldsymbol{K}_t\boldsymbol{\Psi}_t\right)\left(\boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{C}_{t-1}\boldsymbol{B}'_{\theta,t}\boldsymbol{G}'_t + \boldsymbol{Q}_t\right),
\end{aligned}$$

where $\boldsymbol{K}_t = \left(\boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{C}_{t-1}\boldsymbol{B}'_{\theta,t}\boldsymbol{G}'_t + \boldsymbol{Q}_t\right)\boldsymbol{\Psi}'_t\left(\boldsymbol{\Psi}_t\left(\boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{C}_{t-1}\boldsymbol{B}'_{\theta,t}\boldsymbol{G}'_t + \boldsymbol{Q}_t\right)\boldsymbol{\Psi}'_t + \boldsymbol{R}_t\right)^{-1}$. The covariance matrix $\boldsymbol{Q}_t$ can be estimated using discount factors. This involves setting $\boldsymbol{Q}_t$ equal to $\frac{1-\delta}{\delta}\boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{C}_{t-1}\boldsymbol{B}'_{\theta,t}\boldsymbol{G}'_t$, for some fixed value for δ. More details are provided in Prado and West (2010). Another option is to parametrize $\boldsymbol{W}_t$ as $\tau^2 V_t$ where V_t is a spatial covariance matrix. If the prior for τ^2 is inverse Gamma with parameters α_τ and β_τ then the

conditional posterior is

$$\tau^2|\cdot \sim IG\left(\alpha_\tau + \frac{KT}{2}, \beta_\tau + \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{a}_{t-1})'\boldsymbol{V}^{-1}(\boldsymbol{a}_t - \boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{a}_{t-1})\right). \quad (2.12)$$

Given these covariances and the states, the parameters in the kernel and the variance σ^2 can be updated using Metropolis-Hastings steps. The distribution of $\boldsymbol{Y}_t|\boldsymbol{D}_t, \boldsymbol{\theta}, \sigma^2$ is $N(\boldsymbol{\Psi}_t\boldsymbol{m}_t, \boldsymbol{\Psi}_t\boldsymbol{C}_t\boldsymbol{\Psi}_t' + \boldsymbol{R}_t)$. New values for $\boldsymbol{\theta}^*$ are proposed from a proposal distribution $q(\cdot)$. Typically, the variables are transformed so that the proposal distribution could be a normal distribution. The variance of this normal proposal distribution was tuned to an appropriate acceptance rate. If the value of the parameters at the previous iteration of the MCMC is $\boldsymbol{\theta}^{(B-1)}$, then the new values will be accepted with probability

$$\min\left(\frac{p(\boldsymbol{Y}_t|\boldsymbol{D}_t, \boldsymbol{\theta}^*, \sigma^2)p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(B-1)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{Y}_t|\boldsymbol{D}_t, \boldsymbol{\theta}^{(B-1)}, \sigma^2)p(\boldsymbol{\theta}^{(B-1)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(B-1)})}, 1\right)$$

and, otherwise, set $\boldsymbol{\theta}^{(B)} = \boldsymbol{\theta}^{(B-1)}$.

2.5 Summary

We have reviewed some general theory regarding IDE modeling of spatio-temporal data. Additionally, we have provided some motivation for more flexible kernels. Careful study of these principles will result in successful IDE modeling of spatio-temporal data.

Chapter 3

IDE Modeling Using Flexible Parametric Kernels

Equations (2.3) and (2.4) imply that choosing kernels which have more control of higher order moments and tail behavior will allow an IDE to model a broader range of physical processes. There are approaches to define a kernel in an IDE to have full flexibility in moments and tail behavior. One of these is discussed in chapter 4. However, learning the kernel in this setting is very difficult and for some situations, it may be reasonable to assume a simpler model. This chapter explores alternatives to the Gaussian kernel which add one or two parameters, and which are infinitely divisible to match the theory from the previous chapter. The added complexity of the model, therefore, is not too extreme. We introduce these kernels and then explore the advantages of using them instead of a Gaussian kernel. We restrict our attention to one-dimensional space and to kernel parameters which are constant in space. The extensions are more applicable, but for this illustrative setting,

these assumptions will be sufficient to show the advantages of non-Gaussian kernels.

3.1 Alternatives

We present two alternatives to the Gaussian kernel. The first is the asymmetric Laplace, which has an additional parameter which controls skewness. The second is the stable family of distributions which has a parameter controlling skewness and one controlling tail behavior. Both of these kernels are infinitely divisible, which makes the PDE approximations developed in Chapter 2 applicable.

3.1.1 Asymmetric Laplace

The asymmetric Laplace is an infinitely divisible, location family distribution which allows for skewness and heavier tails than the normal. The distribution is characterized by its mode ξ, a scale parameter σ, and a parameter controlling the skewness and other shape properties, $\kappa > 0$. The density function is given by

$$k(x|\xi,\sigma,\kappa) = \frac{\sqrt{2}}{\sigma}\frac{\kappa}{1+\kappa^2}\begin{cases}\exp\left(-\frac{\sqrt{2}\kappa}{\sigma}|x-\xi|\right) & \text{if } x \geq \xi \\ \exp\left(-\frac{\sqrt{2}}{\sigma\kappa}|x-\xi|\right) & \text{if } x < \xi,\end{cases}$$

which shows how the asymmetric Laplace can be formed from two exponentials with different intensities. When $\kappa = 1$, the distribution simplifies to the (symmetric) Laplace distribution. The property of infinite divisibility can be found in Kotz et al. (2001). Figure 3.1 shows different asymmetric Laplace densities for varying values of κ.

The asymmetric Laplace can be written as a mixture of normals with mean $\xi + \mu W$

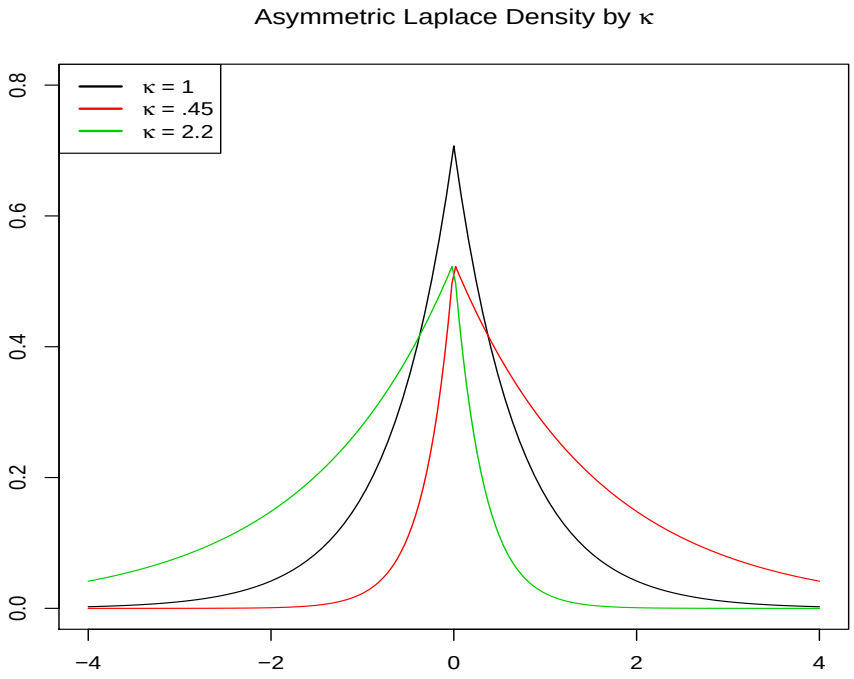


Figure 3.1: Asymmetric Laplace densities for different values of the skewness parameter κ. The distribution is symmetric when $\kappa = 1$ and can be highly skewed in either direction when κ is large or small.

and variance $\sigma^2 W$, where $\mu = 2^{-1/2}\sigma\left(\kappa^{-1} - \kappa\right)$ and W is an exponential distributed random variable with mean 1 (Kotz et al., 2001). This mixture representation yields the following result.

Lemma 4. *The Fourier coefficients of the basis expansion for a kernel in the asymmetric Laplace family are:*

$$b_{2j-1}(s, \boldsymbol{\theta}) = \frac{(1 + .5\rho_j^2\sigma^2)\cos(\rho_j(s+\xi)) + \rho_j\mu\sin(\rho_j(s+\xi))}{(-1 - .5\rho_j^2\sigma^2)^2 + (\rho_j\mu)^2}$$

and

$$b_{2j}(s, \boldsymbol{\theta}) = \frac{(1 + .5\rho_j^2\sigma^2)\sin(\rho_j(s+\xi)) - \rho_j\mu\cos(\rho_j(s+\xi))}{(-1 - .5\rho_j^2\sigma^2)^2 + (\rho_j\mu)^2} \quad .$$

Proof. The asymmetric Laplace distribution can be written as a mixture of normals. If X is a standard normal and W is a standard exponential, then $Y = \xi + \mu W + \sigma\sqrt{W}X$ has an asymmetric Laplace distribution. Hence, conditional on W, Y has a normal distribution with mean $\xi + \mu W$ and variance $\sigma^2 W$. Recall that the $N(\mu, \sigma^2)$ kernel can be decomposed into $\sum_j b_j(s, \boldsymbol{\theta})\phi_j(u)$, where the basis functions are $\phi_{2j-1}(u) = \cos(\rho_j u)$ and $\phi_{2j}(u) = \sin(\rho_j u)$, and the coefficients are $b_{2j-1}(s, \boldsymbol{\theta}) = \exp\left(-.5\rho_j^2\sigma^2\right)\cos(\rho_j(s+\mu))$ and $b_{2j}(s, \boldsymbol{\theta}) = \exp\left(-.5\rho_j^2\sigma^2\right)\sin(\rho_j(s+\mu))$. Therefore, by mixing on W, the asymmetric Laplace distribution coefficients for the Fourier basis expansion can be found through

$$\begin{aligned} b_{2j-1}(s, \boldsymbol{\theta}) &= \int_0^\infty \exp\left(-.5\rho_j^2\sigma^2 W\right)\cos(\rho_j(s + \xi + \mu W))\exp\left(-W\right)dW \\ &= \frac{1}{(-1 - .5\rho_j^2\sigma^2)^2 + (\rho_j\mu)^2}\left[(1 + .5\rho_j^2\sigma^2)\cos(\rho_j(s+\xi)) + \rho_j\mu\sin(\rho_j(s+\xi))\right]. \end{aligned}$$

Similarly, we obtain

$$b_{2j}(s, \boldsymbol{\theta}) = \frac{1}{(-1 - .5\rho_j^2\sigma^2)^2 + (\rho_j\mu)^2}\left[(1 + .5\rho_j^2\sigma^2)\sin(\rho_j(s+\xi)) - \rho_j\mu\cos(\rho_j(s+\xi))\right].$$

□

Computationally, the non-differentiability of the density at its mode makes it harder to approximate using basis functions. To get a working approximation using a Fourier basis, the truncation point required is much larger for the asymmetric Laplace than it is for the Gaussian density, typically ranging from 30 to 100 basis functions. The more skewed the distribution is, the harder it becomes to approximate well.

3.1.2 Stable Distributions

Lemma 3 suggests that the kernel tail behavior will affect IDE evolution. To explore infinitely divisible kernels with tails that are substantially heavier than those of a Gaussian, we consider the family of stable distributions. A distribution belongs to the class of stable distributions if any linear combination of two random variables from a particular class of distributions also belong to that same family. Thus, this is a subset of infinitely divisible distributions, as shown in Samorodnitsky and Taqqu (1997) and Nolan (2003). Stable distributions include the Gaussian, Cauchy, and Levy distributions as special cases shown in Table 3.1. They are governed by 4 parameters, $\mu \in \mathbb{R}$, $c > 0$, $\alpha \in (0, 2]$, and $\beta \in [-1, 1]$, and a wide range of skewness and tail behavior can be achieved by varying the parameters appropriately. A characteristic of the family of stable distributions is that, in general, it does not have an analytically available form for the density function, or moments. These do exist for special cases, such as the Gaussian distribution. The family is generally defined through its characteristic function, which for $\alpha \neq 1$ is given by

$$g(t|\mu, c, \alpha, \beta) = \exp\left\{it\mu - |ct|^{\alpha}(1 - i\beta sgn(t)\tan(\pi\alpha/2))\right\}. \tag{3.1}$$

Figure 3.2 shows how the shape of the density changes with α and β. Note that α controls the tails and β controls the skewness, while μ and c are location and scale parameters, respectively.

Distribution	α	β
Gaussian	2	$[-1, 1]$
Cauchy	1	0
Levy	1/2	1

Table 3.1: Special cases of the stable family of distributions, including the Gaussian, Cauchy, and Levy distributions.

Lemma 5. *The Fourier coefficients of the basis expansion for a kernel in the family of stable distributions are:*

$$\begin{aligned} b_{2j-1}(s, \boldsymbol{\theta}) &= \cos\left(\rho_j(s+\mu) + |c\rho_j|^\alpha \beta sgn(\rho_j)\tan(\pi\alpha/2)\right)\exp\left(-|c\rho_j|^\alpha\right) \\ b_{2j}(s, \boldsymbol{\theta}) &= \sin\left(\rho_j(s+\mu) + |c\rho_j|^\alpha \beta sgn(\rho_j)\tan(\pi\alpha/2)\right)\exp\left(-|c\rho_j|^\alpha\right). \end{aligned}$$

Proof. The stable family of distributions with $\alpha \neq 1$ has characteristic function of the form $g(t) = \exp\left\{it\mu - |ct|^\alpha\left(1 - i\beta sgn(t)\tan(\pi\alpha/2)\right)\right\}$. Decomposing the characteristic function into its real and imaginary parts and then applying Euler's formula, we find the coefficients for the sine and cosine basis functions:

$$\begin{aligned} g(t) &= \cos\left(t\mu + |ct|^\alpha \beta sgn(t)\tan(\pi\alpha/2)\right)\exp(-|ct|^\alpha) \\ &+ i\sin\left(t\mu + |ct|^\alpha \beta sgn(t)\tan(\pi\alpha/2)\right)\exp(-|ct|^\alpha). \end{aligned}$$

The real part of this equation corresponds to the cosine coefficients and the sine part refers to the sine coefficients in a Fourier transform. □

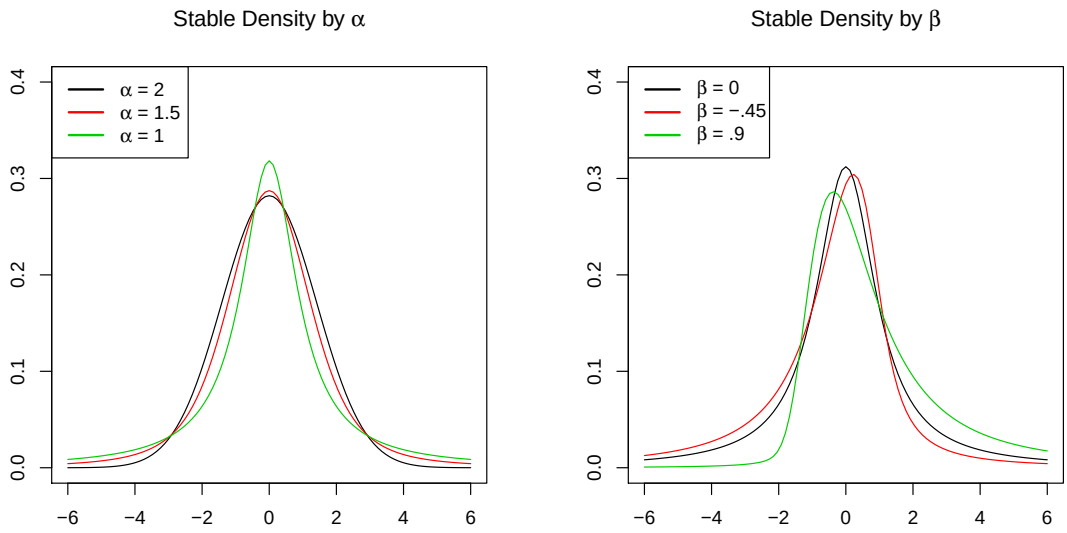


Figure 3.2: Densities from the stable class of distributions with $\mu = 0$ and $c = 1$, for different values of the stability parameter α (left panel) and skewness parameter β (right panel). Smaller α values result in heavier tails and β values far from 0 result in greater skewness. The left panel fixes $\beta = 0$ and the right panel fixes $\alpha = 1.1$.

The quality of this Fourier series approximation depends on the shape of the distribution. To avoid requiring a large truncation point, it is computationally convenient to restrict $\alpha \in (1, 2]$. This restricts the tail behavior to be between the Cauchy and the Gaussian distributions, but still ensures polynomial tail behavior for all values of $\alpha < 2$. For $\alpha < 1$, the required truncation level for a Fourier basis expansion increases tremendously. With $\alpha > 1$, the number of terms required is comparable to the normal, and thus computational expense will be similar. The high degree of flexibility in modeling the heaviness of the tails and the skewness combined with similar computational burden as the Gaussian kernel IDE makes the stable family a very attractive choice for the IDE kernel.

3.1.3 Prior Simulation

To empirically study how the various kernels affect the IDE model, we perform a series of prior simulations under four different kernels. The first of these kernels is normal with a mean of $-.67$ and a variance of 2. The second kernel is an asymmetric Laplace with the same mean and variance as the normal kernel, but with a left skewed density. By matching the means and the variances of these two kernels, we can explore whether the first two moments dominate the IDE process or if a non-zero third moment results in different process realizations, as suggested by equation (2.3). The third kernel is a stable distribution which is skewed and shaped to match the asymmetric Laplace. The final kernel is also a stable distribution with quartiles and a median which match the normal kernel, having heavy tails and no skewness. This will test how tail behavior affects the IDE model for otherwise similar kernels.

Process realizations are simulated from the IDE model according to equation (1.3).

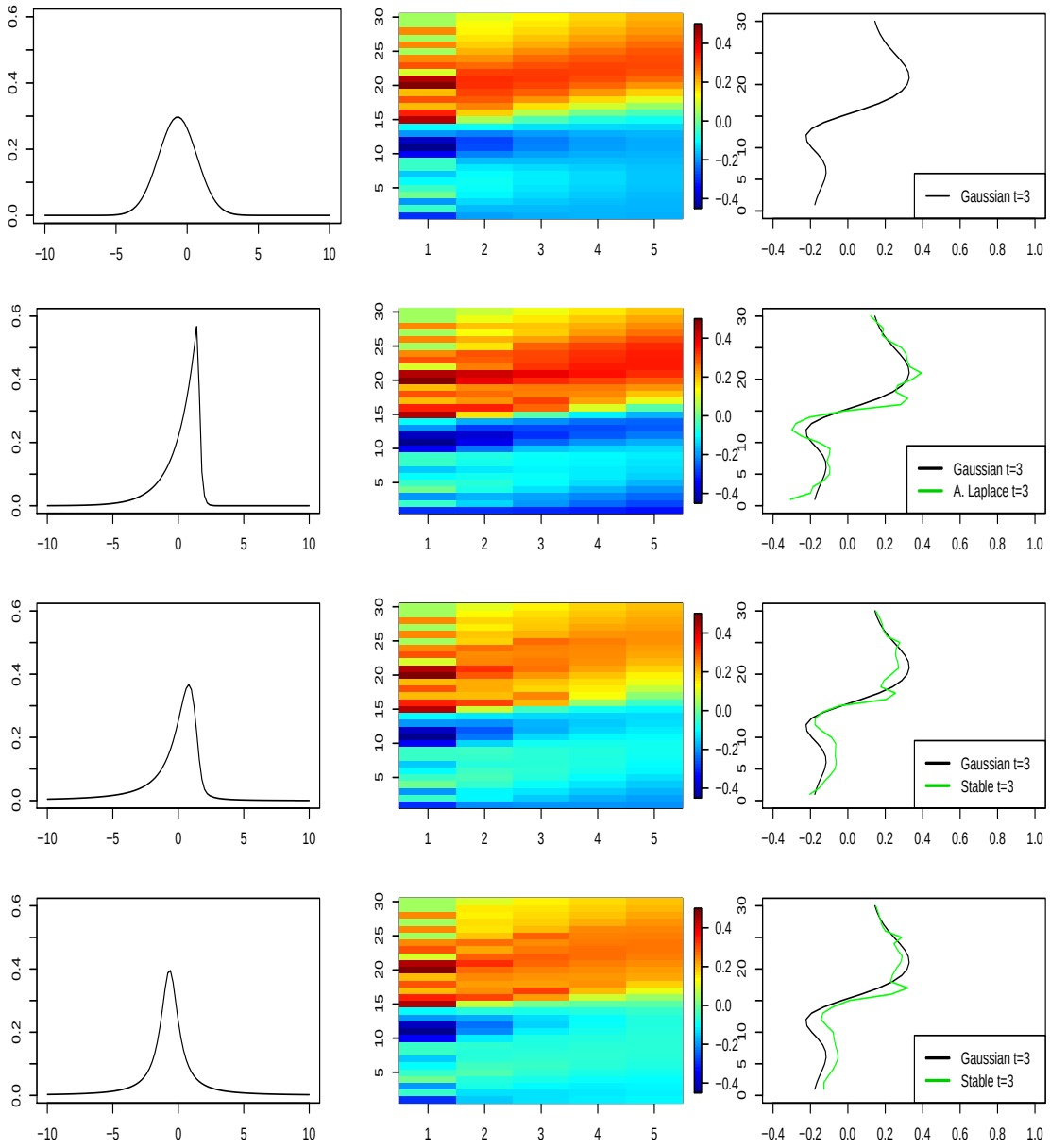


Figure 3.3: IDE prior simulations in one-dimensional space for four distinct kernels. From top to bottom the kernels are normal, asymmetric Laplace, stable with skewness, and stable with heavy tails and no skewness. The first column is the density of the IDE kernel distribution, the second column is the simulated process for 5 time points, and the last column compares the spatial field between the particular kernel and the Guassian kernel for the third time point.

The initial condition at $t = 0$ is a realization from a Gaussian process and the error process is not included, such that the process evolves without any noise added. The resulting IDE processes are shown in Figure 3.3. These simulations show that, while the general trend is similar across each kernel choice, the localized features differ for each time point. The process for the IDE with more flexible kernels behaves as a more colorful version of the process using a Gaussian kernel, as can be seen best in the third column.

3.2 Illustrative Data Examples

The theory supports the use of the asymmetric Laplace and the stable family as possible extensions to the normal distribution for the IDE kernel. To see how these kernels compare in actual model performance, we apply the IDE model with all three kernels and compare the predictive results. In Section 3.2.1 two synthetic data sets will be fit and compared. In Section 3.2.2, the methods will be compared using real data collected by ozonesonde readings on ozone pressure.

3.2.1 Comparing Model Fits with Synthetic Data

To test the asymmetric Laplace and stable distributions against the normal, data is simulated under the IDE setting from two different kernels. The first is a mixture of normal distributions, $.35N(-3, 1) + .25N(-1, 1) + .15N(1, 1) + .1N(3, 1) + .1N(5, 1) + .05N(7, 1)$, which results in a skewed density with exponential tails. The second simulation is from an IDE with a Cauchy kernel, which is a special case of the stable with $\alpha = 1$ and $\beta = 0$. Each of these simulated data sets spans over 200 gridded spatial locations and over 50 time points,

and contains a reasonable amount of observational and process error. Posterior mean and interval estimates for the kernel densities are shown in Figure 3.4. Based on these plots, the Gaussian kernel is unsuccessful in recreating the truth. The more flexible kernels more appropriately capture the skewness and tail behavior of the underlying IDE kernel.

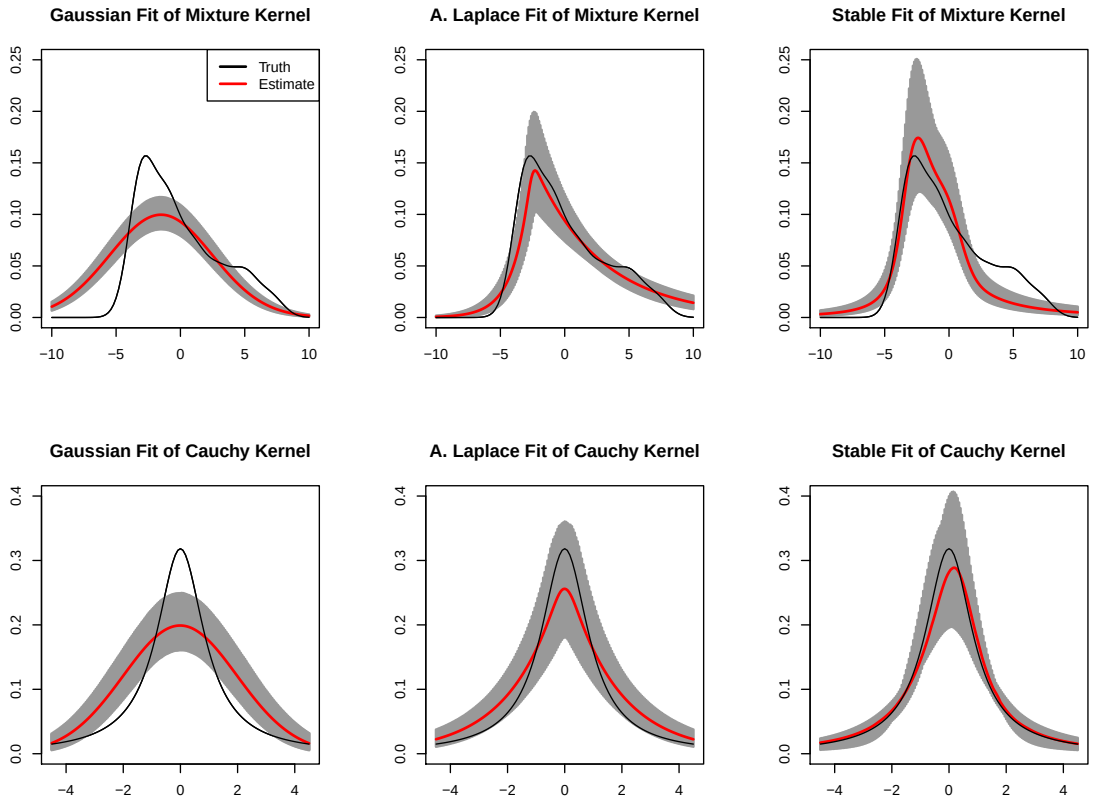


Figure 3.4: Synthetic data. Posterior mean and interval estimates for the IDE kernel density under the model with the Gaussian, asymmetric Laplace, and stable kernels. The top row corresponds to the data generated from an IDE model with a normal mixture for the kernel, and the bottom row to the simulated data based on an IDE model with a Cauchy kernel.

The fitting of the models involves dynamic linear model theory. Because the kernel

parameters are embedded within the structure of the evolution matrix, we use Metropolis-Hastings steps to obtain samples from the posterior distribution of those parameters. Details were given in Chapter 2. We compare the predictions from Markov chain Monte Carlo (MCMC) runs using an energy score, following Gneiting et al. (2008). This procedure allows for simultaneous scoring of a whole spatial field. The energy score is calculated as

$$\hat{es}(F, y) = \frac{1}{m}\sum_{i=1}^{m} ||y^{(i)} - y|| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} ||y^{(i)} - y^{(j)}||, \tag{3.2}$$

where $y^{(1)}, ..., y^{(m)}$ are samples from F, the posterior predictive distribution and y denotes the data vector. For each of the simulated data sets we compute energy scores for one step ahead out-of-sample predictions for 50 time points. Table 3.2 shows the percentage of times each of the kernels scored the lowest.

	True kernel	
Fitted kernel	Mixture	Cauchy
Gaussian	16%	0%
Asymmetric Laplace	70%	4%
Stable	14%	96%

Table 3.2: Synthetic data. The percentage of times for which each of the kernels had the lowest energy score for each of the simulated data sets. The asymmetric Laplace performed the best for the mixture kernel and the stable family performed the best for the Cauchy kernel.

The scoring indicates clearly which kernel performs the best in each case. The asymmetric Laplace outperforms the others for the skewed mixture, whereas the stable distribution outperforms the others for the Cauchy kernel. To offer an explanation, note that the polynomial tails of the stable may not match up well with the exponential tails of the

mixture IDE kernel and the Gaussian could not capture the skewness, but the asymmetric Laplace is able to capture skewness and tail behavior better. Whereas with the Cauchy IDE kernel, only the stable distribution could match the polynomial tails.

3.2.2 Ozone Data

The study of ozone has provided an abundant source of environmental and statistical literature over the past decades. The effect of lower atmosphere ozone measurements has been seen to affect other climate variables such as concentration of certain pollutants and temperature (Robeson and Steyn, 1990). Other studies have shown how ozone concentrations affect crop yields and other agricultural variables (Heck et al., 1984). Understanding lower atmosphere ozone levels may help to understand and predict many other important variables which have a direct societal impact.

To study how the kernel choice may affect IDE model performance, we fit our proposed models to 10 years of low atmosphere ozone pressure data. These data are collected by ozonesondes, which are balloons that ascend into the atmosphere and record measurements at regular intervals. The data set we study includes biweekly ozone pressure from October 1996 to October 2006 collected at Koldewey Station near the North Pole. Details about this weather station and others related to it can be found at http://www.awi.de/en/home/.

The data is collected by releasing a balloon in the air which, at certain intervals throughout its flight, takes a measurement of ozone pressure in millPascals (mPa). The resulting data structure poses many issues for modelers. First, the locations at which the data are collected vary across time. The balloon usually takes measurements at regular intervals, but this rarely corresponds to a consistent pattern with surrounding time points.

Another issue is that the data collecting mechanism would often fail to reach higher altitudes, leaving the entire upper half of the observation interval missing. Yet another major issue is that the data is somewhat irregular and hard to model using standard methods. For this particular illustration, we restrict our focus to the first 6,000 feet, which corresponds to lower-atmosphere ozone pressure. The data is collected almost every week, though several weeks are missing. Since this is an illustration, we opt to use biweekly data to avoid missing time points.

Because the balloon moves only in one direction, the domain for space is one-dimensional. The data is displayed in Figure 3.5 by altitude and time. There are a few stretches with outlying observations which are included in the analysis but are not shown so that the finer details of the data can be viewed, and also to help compare with the fitted values in Figure 3.9. In Figure 3.5, we note a potential seasonal trend. To account for this seasonality, we add two harmonics, $\boldsymbol{Z}_{ti} = (Z_{ti}^{(1)}, Z_{ti}^{(2)})$, for $i = 1, 2$. These variables will evolve through a rotation matrix with frequency λ_i. The resulting process has a cyclical forecast function with a period of $2\pi/\lambda_i$ (West and Harrison, 1997, Chp. 8). By including two harmonics we can account for seasonal variability with two different periods. The full model is

$$
\begin{aligned}
Y_t(s)|X_t(s), Z_{t1}^{(1)}, Z_{t2}^{(1)}, \sigma^2 &= X_t(s) + Z_{t1}^{(1)} + Z_{t2}^{(1)} + \varepsilon_t(s), \quad \varepsilon_t(s) \overset{i.i.d.}{\sim} N(0, \sigma^2) \\
X_t(s)|\{X_{t-1}(s) : s \in D\}, \boldsymbol{\theta} &= \int_D k(s-u|\boldsymbol{\theta}) X_{t-1}(u) du + \omega_t(s) \\
\begin{pmatrix} Z_{ti}^{(1)} \\ Z_{ti}^{(2)} \end{pmatrix} &= \begin{pmatrix} \cos(\lambda_i) & \sin(\lambda_i) \\ -\sin(\lambda_i) & \cos(\lambda_i) \end{pmatrix} \begin{pmatrix} Z_{t-1,i}^{(1)} \\ Z_{t-1,i}^{(2)} \end{pmatrix} + \nu_t, \quad i = 1, 2 \\
\sigma^2 \sim \text{gamma}(a, b), \quad \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}), \quad \nu_t | W_t^Z \sim N_2(0, W_t^Z),
\end{aligned}
$$

where, for any points $s_1, ..., s_n$, the vector $(\omega_t(s_1), ..., \omega_t(s_n))$ has a normal distribution with a zero mean and covariance function W_t. The state variables $\boldsymbol{a}_0, ..., \boldsymbol{a}_T$ are sampled using forward filtering backwards sampling techniques described in section 2.4. The matrices W_t and W_t^Z are modeled using discount factors, which is also discussed in section 2.4. The IDE kernel is chosen to be Gaussian, asymmetric Laplace, and then stable in three different model fits. The prior parameters a and b are fixed. An exploratory analysis was performed where the periods of the two harmonics were included as parameters in the model and a cluster of posterior mass around 6 months and 12 months was observed. The two harmonics were then fixed to have periods of 6 and 12 months, which results in $\lambda_1 = 2\pi/26$ and $\lambda_2 = 2\pi/13$, assuming 52 weeks per year.

For the kernel parameters and the observational variance, the posterior distributions are robust to a wide range of priors. We use a $N(0, 300^2)$ prior for the location parameter, and a gamma$(1, .01)$ prior for the scale parameter in each case. The skewness parameter κ in the asymmetric Laplace received a gamma$(1, 1)$ prior. The stable parameters α and β were assigned scaled Beta$(2, 2)$ prior distributions to match their support. The important prior specification is for $\boldsymbol{m}_0$ and $\boldsymbol{C}_0$, which are the mean and covariance, respectively, of the basis coefficients for the time 0 process. Poor choices for these can greatly affect the posterior for the kernel parameters. Ozone pressure typically does not stray too far from the range of 2 to 4. Our best guess of the time 0 process is a constant function at 3. The basis coefficients that define $\boldsymbol{m}_0$ are $(3/\sqrt{2r}, 0, ..., 0)$, where r is the period of the Fourier transform used for the basis. We constructed $\boldsymbol{C}_0$ as a diagonal matrix with decreasing values down the diagonal so that variances of the higher order terms of

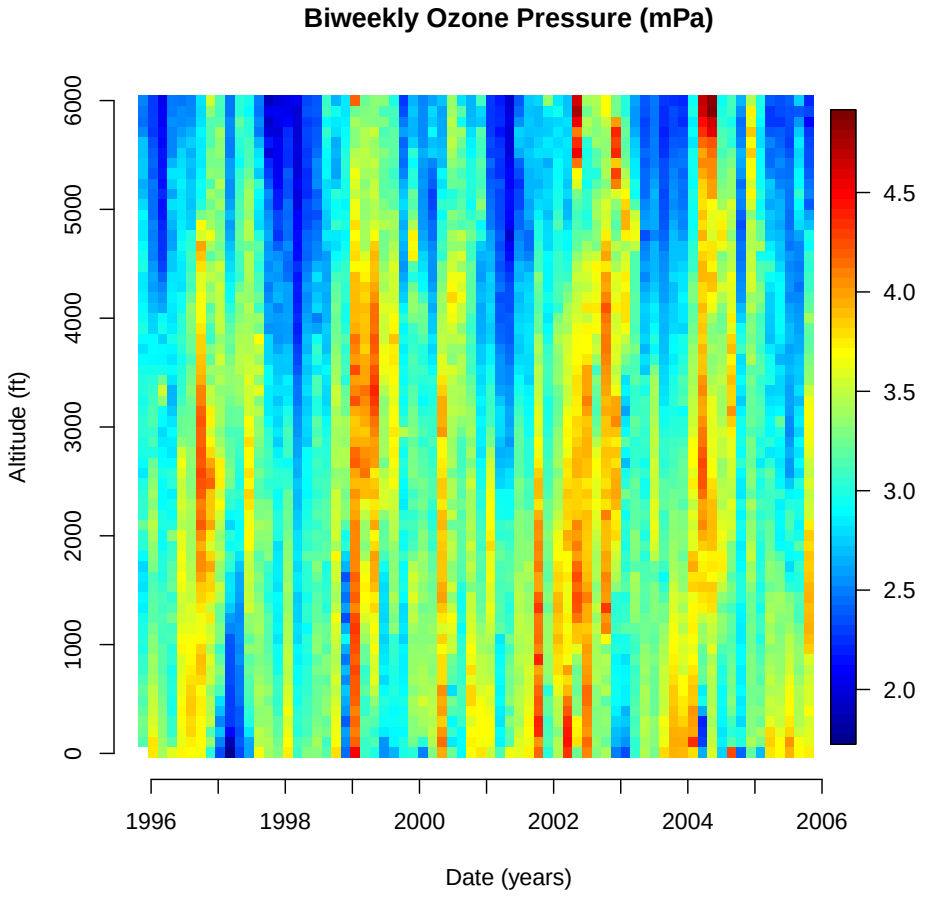


Figure 3.5: Biweekly ozone pressure measured on a vertical profile, plotted across altitude (0 to 6,000 feet) and over time (October 1996 to October 2006).

the basis function expansion are close to 0. We present results based on 50,000 samples for every parameter of every model using MCMC after a burn-in of 50,000. Convergence was confirmed using several methods found in the "boa" package in R (Smith, 2007).

To demonstrate the practical utility of non-Gaussian IDE kernels, we can study the kernel estimates, and the posterior distribution of the parameters which control skewness and heavy tails. The posterior mean estimates for the kernel density under each model are shown in Figure 3.6, and it can be clearly seen that the kernel tends to be asymmetric for models where that is allowed. The posterior distribution for κ in the asymmetric Laplace, and the stable parameters α an β are shown in Figure 3.7. Recall that κ controls the skewness of the asymmetric Laplace, and α and β control the tail behavior and skewness of the stable distribution. The credible intervals for each of these parameters are shown in Table 3.3. The asymmetric Laplace parameter κ includes 1 in the credible interval, suggesting that we can not rule out symmetry based on the parameter estimates. However, the credible interval for the stable distribution parameter β does not include 0, suggesting that the model with the stable distribution kernel is not symmetric. Figure 3.8 shows profiles of the fitted values of one step ahead predictions for three observations from the data set using each kernel. Using such profiles, it can be seen that the Gaussian kernel IDE does not appropriately model ozone pressure in several regions. The stable distribution, however, seems to perform much better. The model residuals for the stable distribution IDE model are shown in Figure 3.9.

The one step ahead predictions shown in Figure 3.8 are calculated for the data set for all three models. As in the simulated example, these predictions can be scored using the

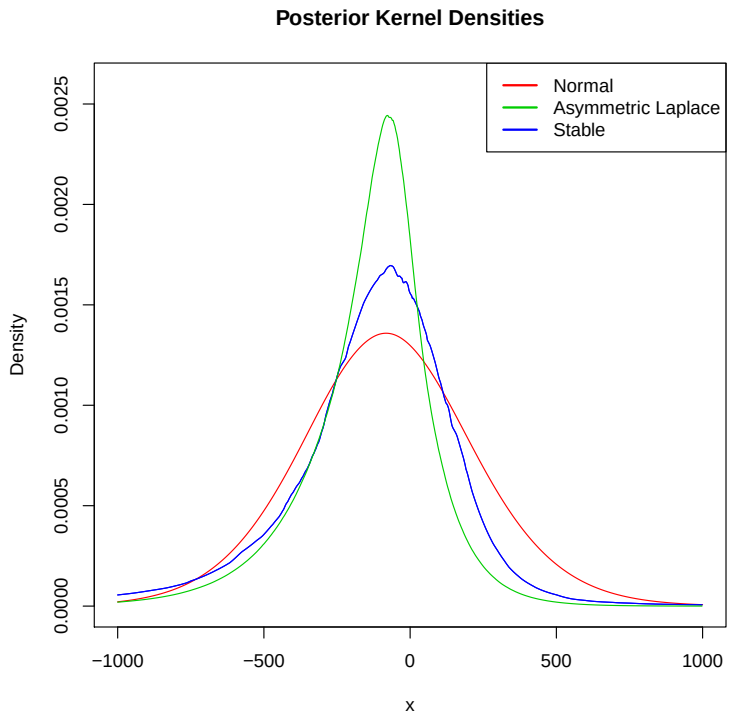


Figure 3.6: Ozone data. Posterior mean estimates for the IDE kernel under the Gaussian, asymmetric Laplace, and stable models.

Parameter	Median	2.5%	97.5%
κ	1.22	.69	1.75
α	1.48	1.26	1.80
β	$-.57$	$-.86$	$-.17$

Table 3.3: Ozone data. Posterior median and 95% credible intervals for certain parameters of the IDE models with non-Gaussian kernels.

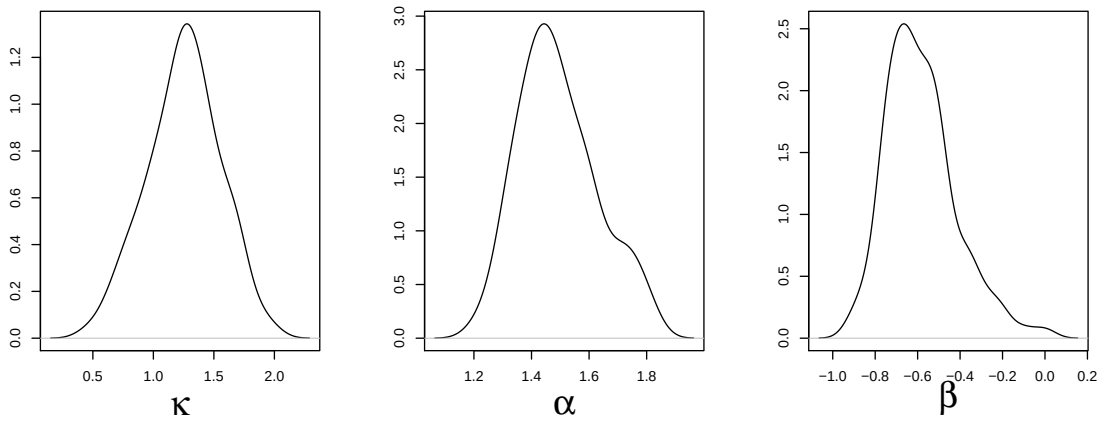


Figure 3.7: Ozone data. Posterior density for the skewness parameter κ of the asymmetric Laplace kernel (left panel). Posterior densities for parameters α and β which control the tails and the skewness of the stable kernel (middle and right panel).

measure in equation (3.2). The results of the energy scores can help determine which model performs the best in one step ahead predictions. Table 3.4 shows each possible ordering for the scores and how often they occur. Recall that lower scores refer to a better fit. The stable distribution has the lowest energy score for 73% of the observations. Only 10% of the observations have the Gaussian kernel as the lowest score. These scores help discern the differences that the figures themselves are not able to show. It is clear that the particular criterion favors the stable distribution over the Gaussian and asymmetric Laplace.

To summarize these results, the posterior distribution of the stable kernel parameters suggest that normality and symmetry are poor assumptions for the IDE kernel. The posterior distribution of the asymmetric Laplace does not rule out symmetry, but the estimate shown in Figure 3.6 for the asymmetric Laplace kernel is clearly asymmetric. Scoring

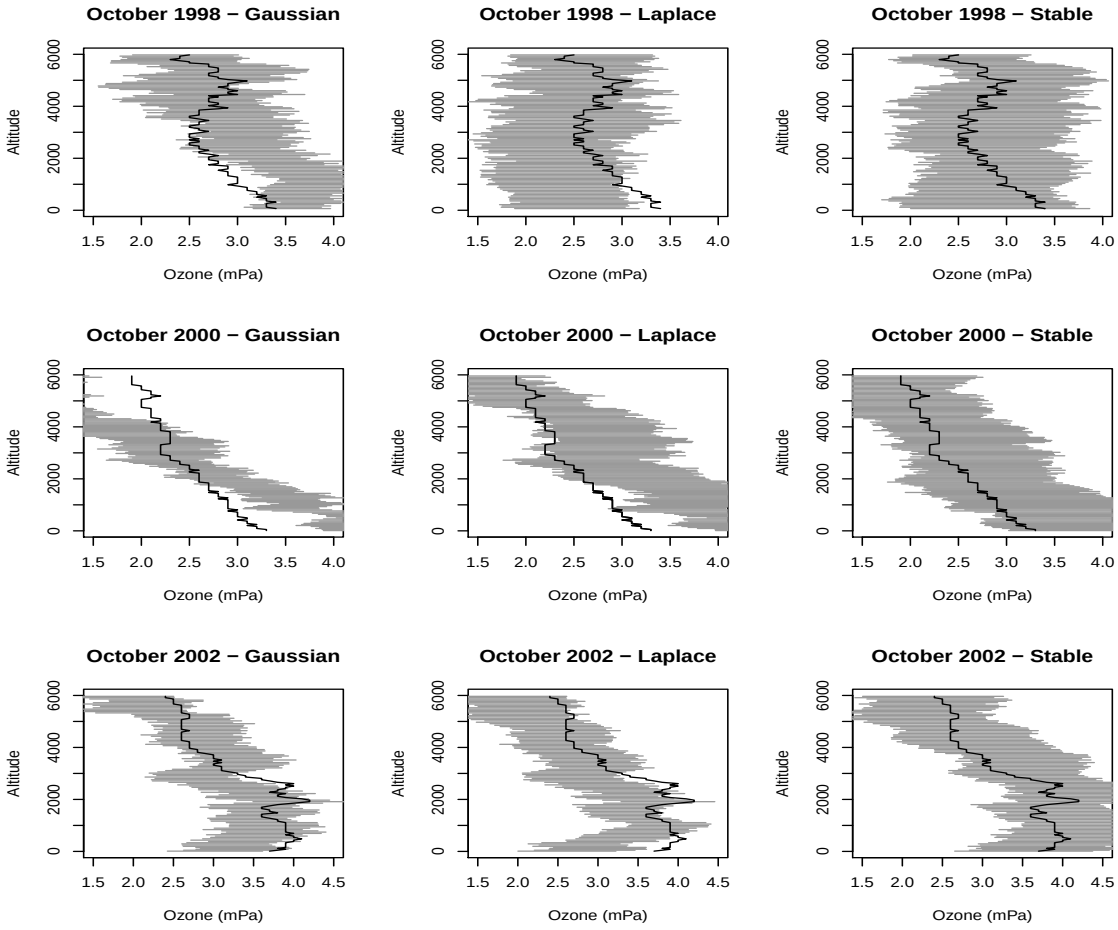


Figure 3.8: Ozone data. The profiles of ozone concentration are shown for three different months with 95% credible intervals shaded in for each of the three different kernels.

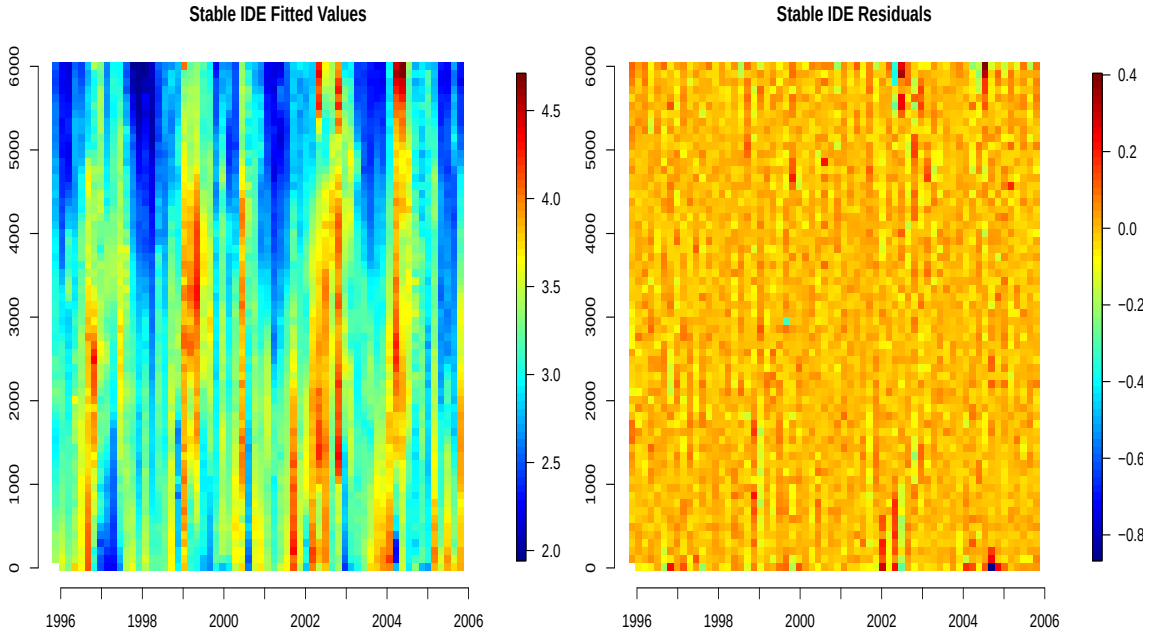


Figure 3.9: Ozone data. Fitted values (left) and residuals (right) are shown for the fitted model with the stable distribution kernel for every observation. The overall fit is good with the exception of a few outlying stretches.

Order	Frequency
S<AL<G	53%
AL<S<G	11%
G<S<AL	6%
S<G<AL	20%
AL<G<S	6%
G<AL<S	4%

Table 3.4: Ozone data. Each possible ordering for the scores of the IDE models under the three distinct kernels are shown with the percentage of observations that the scores followed that order. S refers to the stable distribution, AL to the asymmetric Laplace, and G to the Gaussian kernel.

procedures for the out of sample predictions suggest that the model with the stable kernel distribution performs the best in terms of predictive model accuracy.

3.3 Conclusion

Spatio-temporal data often present complicated space-time interactions that are difficult to model accurately. Under the IDE model framework, electing to use a kernel more flexible than the Gaussian, which is used in nearly all IDE modeling, provides better predictive accuracy and more potential for successfully capturing the spatio-temporal evolution of the field. Compared to Gaussian kernel densities, kernels with flexible tail behavior and potential skewness, facilitate more complicated transfer of dynamics from one time point to the next. In this paper, we have shown how the choice of kernel influences the process through theory, simulations, and data analysis. We have proposed two alternative kernel families with desirable theoretical and computational properties.

Computations for the models proposed in this paper are based on truncated expansions on Fourier bases. In our experience, asymmetric Laplace kernels require a larger number of coefficients than stable distribution kernels, with $\alpha > 1$. The latter can be well approximated with a computational effort comparable to the one needed for Gaussian kernels. Hence, when alternatives to the Gaussian IDE model are needed for very large data sets, the stable family of distributions seems a more practical choice than the asymmetric Laplace.

The models proposed in this paper can be extended in at least three different ways. First, we can place a Gaussian process prior on the location parameter and scale parameters,

along the lines of (Wikle, 2002). This will achieve non-stationarity. The remaining kernel parameters can also be spatially varying by using, for example, transformations of Gaussian processes. Second, we can further extend the flexibility of the kernel shape by considering non-parametric representations of the kernel. This extension is discussed in Chapter 4. The third extension pertains the development of models with non-Gaussian kernels for spaces of dimension higher than one, most importantly, two dimensions. Conceptually, this extension is straightforward. Nevertheless, inference and computations for the families proposed are quite challenging. Starting from a multivariate characteristic function, it is possible to obtain the Fourier basis expansion in order to evaluate the IDE integral, along the lines of the one-dimensional case. For the asymmetric Laplace, a two-dimensional characteristic function is readily available (Kotz et al., 2001). However, the computational burden due to the large number of basis functions required for a good approximation of the kernel is compounded by the dimensionality. For the stable family, multivariate generalizations are not immediate. Bivariate stable kernels are explored in Chapter 5.

Chapter 4

Bayesian Non-parametric Kernel IDE Modeling

4.1 Introduction

The previous chapter focused on simple extensions of the IDE kernel from Gaussian to more flexible families of distributions. The result was an improvement in model accuracy and prediction. Building upon this idea, we propose non-parametric kernels for use in IDE models. This chapter begins with a review of Dirichlet process mixtures in section 4.2, which will be used as a prior for the kernel. The methods used for fitting the data are presented in section 4.3 They are similar to those presented in Chapter 2, however, the parameter set in the kernel increases tremendously. Advanced MCMC methods will be used to learn the kernel parameters. To decrease computational time, Hermite polynomials will be used to construct a basis which may be more effective at approximating the Gaussian density than

more traditional methods. Simulations in section 4.4 will show that the model can learn spatially varying kernels. These methods are applied to the ozone data set in section 4.5 and posterior results are compared between the parametric models studied in Chapter 3 and the non-parametric kernel IDE models presented in this chapter.

4.2 Dirichlet Process and Dirichlet Process Mixtures

A Dirichlet process (DP) is a random probability measure on the space of probability distributions (Ferguson, 1973). The DP is a core modeling tool for Bayesian non-parametric methods, especially after the constructive definition was introduced in Sethuraman (1994). The constructive definition of a Dirichlet process, $G \sim DP(\alpha, G_0)$, is $G = \sum_{l=0}^{\infty} w_l \delta_{\theta_l}$ where each θ_l is drawn *i.i.d.* from the base distribution G_0 for all l. The weights come from a process commonly referred to as stick-breaking, where latent variables $\xi_1, \xi_2, ...$ are drawn *i.i.d.* from a $Beta(1, \alpha)$ distribution, and the weights are assigned as $w_l = \xi_l \prod_{i=1}^{l-1}(1 - \xi_i)$. The parameter α controls how close the random distributions will be to the base distribution G_0, with larger values of α leading to realizations which are closer to G_0. To define a flexible continuous kernel, the parameters of a Gaussian distribution is mixed with a DP (Antoniak et al., 1974). The result is a very flexible continuous distribution which can capture heavy tails, light tails, skewness of any kind, and complete flexibility of higher order moments. Because the weights of the constructive definition of the DP are generally decreasing, the constructive definition can be reasonably truncated to a finite sum. The choice for the number of weights to use depends on α and the data, but it should be considered carefully. When the DP is truncated, the final weight, w_L, is equal

to $\prod_{l=1}^{L-1}(1-\xi_l)$, and should be very small.

The construction of spatially varying kernels in the IDE literature has focused mainly on the Gaussian kernel. A more flexible prior model for spatially varying kernels would be a spatial Dirichlet process (SDP) mixture (Gelfand et al., 2005). An SDP is a specific application of dependent Dirichlet processes defined by MacEachern (2000). As illustrated in Kottas et al. (2008), spatial and spatio-temporal data has been analyzed using an SDP mixture of normals model. Though we will work with the basic version, there are generalizations, such as the ones found in Duan et al. (2007). Using the constructive definition, a spatial DP can be formulated similar to that of the DP. The weights are still found using stick breaking, but the atoms are now realizations from a spatially dependent process, $G_0(s)$. The resulting random measure is $G(s) = \sum_{l=1}^{L} w_l \delta_{\theta_{l,D}}$ where $\boldsymbol{\theta}_{l,D} = \{\theta_l(s) : s \in D\}$. For any finite set of points $s_1, ..., s_n$, the atom, $\theta_l(s_1), ..., \theta_l(s_n)$, forms a vector drawn from a multivariate normal, arising as the finite dimensional distribution of $G_0(s)$. Flexible continuous kernels will again be obtained by mixing the mean of a normal kernel with the SDP.

By using an SDP mixture for the kernel, we are allowing the kernel to change across the entire spatial field. The power of this model is that it can capture drastically different kernel behaviors for different locations. For example, when using an IDE model with an SDP mixture kernel, one region may exhibit long-tail dependence while another exhibits thinner tails. Neither stationary DP kernels nor spatially varying Gaussian kernels will allow for such characteristics to be modeled.

4.3 Methods

Using an appropriate truncation, the kernel from equation (2.2) for an IDE with a DP mixture kernel is $k(u|s, \boldsymbol{\theta}) = \sum_{l=1}^{L} w_l \phi(u|\mu_l, \sigma^2)$, where $\phi(\cdot)$ is the density corresponding to a normal distribution. Each μ_l is drawn *i.i.d.* from a $N(\mu_0, \sigma_0^2)$ base distribution and the weights arise from stick breaking. Equation (2.2) can be written for this model as

$$\begin{aligned} X_t(s) &= \int \sum_{l=1}^{L} w_l \phi(u; s + \mu_l, \sigma^2) X_{t-1}(u) du + \eta_t(s) \\ &= \sum_{l=1}^{L} w_l \int \phi(u; s + \mu_l, \sigma^2) X_{t-1}(u) du + \eta_t(s), \end{aligned}$$

which is a weighted sum of L Gaussian kernel IDE models with different means.

When the kernel is an SDP mixture of normals, μ_l is replaced by $\mu_l(s)$, where, for any set D, $\{\mu_l(s) : s \in D\} \sim G_{0,D}(\mu_0(s), \Sigma_0(s, s'))$, where $\mu_0(s)$ is a mean function and $\Sigma_0(s, s')$ is a covariance function corresponding to the Gaussian process $G_{0,D}$. Typically, the process will be evaluated at a finite number of locations, $s_1, ..., s_n$, so the base distribution is defined through the finite dimensional distribution of the underlying Gaussian process for any set of points $s_1, ..., s_n$. In both the DP mixture and the SDP mixture kernel IDE models, the result that the process is a weighted sum of Gaussian kernel IDE models will be used to find an appropriate method for fitting the model. Each Gaussian kernel IDE component will be approximated individually using a basis function decomposition and then recombined using a weighted sum.

Using these representations, the hierarchical IDE model from equations (2.8) and

(2.9) with a Dirichlet process mixture kernel can be written as

$$\begin{aligned} Y_t(s) &= \sum_{i=1}^{K} a_i(t)\psi_i(s) + \epsilon_t(s) \\ \sum_{i=1}^{K} a_i(t)\psi_i(s) &= \sum_{l=1}^{L}\sum_{i=1}^{K} a_i(t-1)b_i(s,\boldsymbol{\theta}_l) + \eta_t(s), \end{aligned}$$

where $b_i(s,\boldsymbol{\theta}_l)$ is the i-th basis coefficient of the basis expansion of the l-th component of the mixture.

4.3.1 Hermite Polynomial Basis

As discussed in section 2.3.1, the number of basis functions used for the series expansion approximations is directly related to the accuracy of the approximation to the kernel and inversely related to the computational speed. A poor kernel approximation invalidates any attempts to interpret physical characteristics of the process implied by the kernel. For example, Xu et al. (2005) connects the mean of a spatially varying Gaussian kernel to average wind speed. Connections such as these would be impossible if the kernel is not accurately represented by the basis. Conversely, relying on a large number of basis functions will improve accuracy, but may make practical computation unreasonable. The choice of which basis function to use and how many to include are made prior to fitting the model. Section 2.3.1 also discussed practical usage of two popular basis function choices, the Fourier basis and Empirical Orthogonal Functions (EOFs). The advantage of the Fourier basis is that $b_i(s,\boldsymbol{\theta})$ is related directly to the characteristic function for any distribution used as the kernel. However, a large spatial range compared to the width of the kernel can lead to poor approximations or will require a very large number of basis functions. EOFs are a popular choice for dimension reduction of spatial and space-time models (Cressie and

Wikle, 2011). This may be an excellent choice for representing the process, except it is a discrete basis and requires a regular grid. Also, while it approximates the process well, it typically requires a large truncation to approximate the kernel with desired accuracy. The solution we present is the orthonormal basis corresponding to Hermite polynomials (Olver, 2010).

Physicist's Hermite polynomials are defined as $H_n(x) = \left(2x - \frac{d}{dx}\right)^n \cdot 1$. These polynomials are orthogonal with respect to the weight function $w(x) = e^{-x^2}$. The specific inner product for Hermite polynomials is $\int H_m(x)H_n(x)w(x) = \sqrt{\pi}2^n n!\delta_{nm}$. If we define new polynomials $h(x) = H(x)\sqrt{w(x)}/\sqrt{(\sqrt{\pi}2^n n!)}$ then $h(x)$ forms an orthonormal basis, meaning that $\int h_m(x)h_n(x) = \delta_{nm}$. These functions, $h_0,\ h_1,\ h_2, ...,$ are called Hermite functions and they provide a basis which can be used for the series expansion in IDE modeling. When using Hermite functions in a series expansion approximating a normal density with mean μ and variance σ^2, the coefficient corresponding to the n-th Hermite function is

$$b_n(\mu, \sigma^2) = \frac{1}{\sqrt{(\sqrt{\pi}2^n n!)\,(1+\sigma^2)}} \exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right) \sum_{k=0}^{n} H_{n,k} m_k \tag{4.1}$$

where $H_{n,k}$ is the k-th coefficient in the n-th Hermite polynomial and m_k is the k-th raw moment of a normal distribution with mean $\mu/(\sigma^2+1)$ and variance $\sigma^2/(\sigma^2+1)$. Then decomposing the kernel results in $k(u|s, \boldsymbol{\theta}) = \sum_{n=0}^{K} b_n(\mu, \sigma)h_n(x)$ and decomposing the process yields $X_t(s) = \sum_{m=0}^{K} a_m(t)h_m(x)$. Because of the orthogonality of the Hermite functions, the IDE integral in equation (2.2) becomes $\sum_{n=1}^{K} a_n(t-1)b_n(s, \boldsymbol{\theta})$, exactly what is needed to obtain the representations in equation (2.5).

Figure 4.1 shows the shape of a few of these Hermite functions, as well as illustrates

a potential problem. As can be seen, the squared exponential weight function causes the

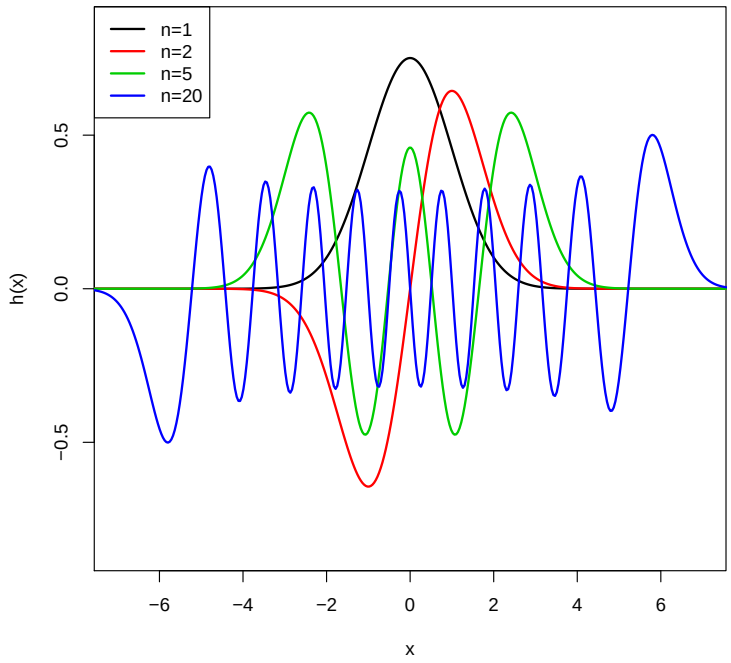


Figure 4.1: A few of the Hermite function basis functions are shown. The relative range of the function increases as n increases.

function to decrease to 0 far from the origin, so Hermite functions are limited past a certain value. For example, if the first 20 Hermite functions are used to estimate a kernel centered at 15 and a variance of 1, the approximation would be inappropriate because the Hermite functions near the mode of the distribution are restricted by the weight function. To avoid this, the spatial locations for the data should be scaled. Resulting estimates of kernel densities can be scaled back without consequence after they are sampled. If the spatial locations are $s_1, s_2, ..., s_n$ with a range of $s_n - s_1 = R$ and a center of c, define new spatial locations $s_1^*, ..., s_n^*$ where $s_i^* = (s_i - c)/(R/4)$. These new spatial locations are restricted

Number of basis functions	Suggested range
10	(-4,4)
20	(-5,5)
30	(-6.5,6.5)
40	(-8,8)
50	(-9,9)

Table 4.1: Suggested ranges are given for the corresponding number of basis functions.

between -4 and 4, but relative distances between adjacent points remain the same. The actual distances will change, however, requiring such things as the range parameters in the process covariance structure to be adjusted. Scale back the spatial locations afterwards by $s_i = (R/4)s_i^* + c$. If the new range is too small with respect to the number of basis functions chosen, there is a risk of computational singularities when sampling from the posterior. If the new range is too large, the approximation to the kernel is potentially inaccurate. Some suggested ranges are given in Table 4.1. These ranges are found by graphical exploration using a kernel with a standard deviation of .25. Using more than 60 basis functions may require changing the floating point precision due to automatic rounding of small coefficients, which can significantly alter the higher order terms of the polynomial. However, 50-60 Hermite basis functions will typically be more than sufficient for an accurate representation of kernel shapes within a range of about -9 to 9. Depending on how wide the range is with respect to the estimated kernel, a Hermite function basis may accurately approximate a mixture of normals kernel with as few as 10 basis functions. Figure 4.2 shows how the Hermite basis and Fourier basis differ when approximating the normal density using a small number of basis functions. The difference can become even more dramatic when the kernel width becomes smaller. There may be many reasons the Hermite basis performs better than

the Fourier basis when approximating the Gaussian kernel. The first is that the exponential weight of the Hermite functions matches the exponential tails of the Gaussian kernel, while the Fourier basis is composed of functions with sinusoidal tails. Also, the range used for the Fourier basis needs to be expanded past the data to account for its periodic nature. Otherwise, the kernel would wrap around the edge of the data giving weight to the process at locations on the opposite side of the range. The Hermite functions do not share this concern.

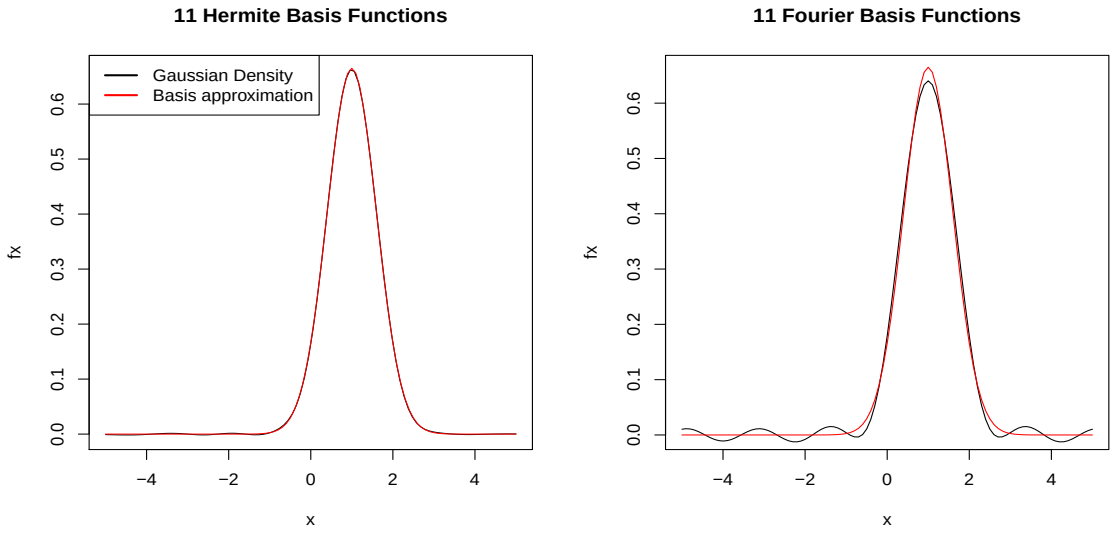


Figure 4.2: The approximation to a normal density with mean 1 and variance $.6^2$ is compared using 11 Hermite basis functions and 11 Fourier basis function over a range of -4 to 4.

4.3.2 Posterior Inference

The model to learn is given in equations (2.8) - (2.10). Section 2.4 outlines exactly how to use FFBS to sample from the state parameters conditional on the kernel parameters and observational and process variance. They can then be used in a Gibb's sampler to

conditionally sample from the posteriors of the other parameters. Standard kernel parameter estimation involves Metropolis-Hastings. However, there may not be information in the likelihood about the kernel parameters when using a DP mixture kernel to correctly learn the posterior kernel. Estimation for DP kernel parameters is typically done by clustering data (Ishwaran and James, 2001). Since no data is available to cluster and standard MCMC methods will not work, we proposed advanced MCMC methods to learn the non-parametric kernel.

Hamiltonian Markov Chain Monte Carlo (HMCMC) introduces latent variables, $p_1, ..., p_m$ for each parameter. These represent the momentum of the path the parameter follows in the HMCMC chain and have $N(0, M_i)$ priors. The momentum and position of the proposal will change according to Hamiltonian physics. Standard Metropolis-Hastings may take an excessive amount of iterations to converge to the posterior, if it ever does. HMCMC is better suited for the large number of parameters and the complicated way they are embedded in the likelihood. Not every parameter needs to be estimated using HMCMC. For example, the atoms of the DP mixture may be sampled using HMCMC while the common variance of the mixture component kernels may be estimated using standard MCMC.

Hamiltonian Monte Carlo

Hamiltonian Markov chain Monte Carlo (HMCMC) involves taking the derivative with respect to unknown parameters of the negative log of the target function, which in this case is the posterior (Neal, 2011). If the prior of the parameter the derivative is being

taken with respect to is $p(\theta)$, then the negative log of the relevant parts of the posterior is

$$-l(\theta) = -\log(p(\theta)) + \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{a}_t - \boldsymbol{\Psi}'\boldsymbol{B}_\theta\boldsymbol{a}_{t-1}\right)'\boldsymbol{W}_t^{-1}\left(\boldsymbol{a}_t - \boldsymbol{\Psi}'\boldsymbol{B}_\theta\boldsymbol{a}_{t-1}\right).$$

Expanding this out results in

$$-l(\theta) = -\log(p(\theta)) + \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{a}_t'\boldsymbol{W}_t^{-1}\boldsymbol{a}_t - 2\boldsymbol{a}_t\boldsymbol{W}_t^{-1}\boldsymbol{\Psi}'\boldsymbol{B}_\theta\boldsymbol{a}_{t-1} + \boldsymbol{a}_{t-1}'\boldsymbol{B}_\theta'\boldsymbol{\Psi}\boldsymbol{W}_t^{-1}\boldsymbol{\Psi}'\boldsymbol{B}_\theta\boldsymbol{a}_{t-1}\right)$$

Using a handful of properties from matrix calculus we can find

$$\frac{d(-l(\theta))}{d\theta} = -\frac{d(-\log(p(\theta)))}{d\theta} + \frac{1}{2}\sum_{t=1}^{T} -2\boldsymbol{a}_t\boldsymbol{W}_t^{-1}\Phi'\frac{d\boldsymbol{B}_\theta}{d\theta_i}\boldsymbol{a}_{t-1} + 2\,tr\left(\boldsymbol{a}_{t-1}\boldsymbol{a}_{t-1}'\boldsymbol{B}_\theta\boldsymbol{\Psi}\boldsymbol{W}_t^{-1}\boldsymbol{\Psi}'\frac{d\boldsymbol{B}_\theta}{d\theta_i}\right),$$

where $\frac{d\boldsymbol{B}_\theta}{d\theta}$ is a element-wise derivative of $\boldsymbol{B}_\theta$ with respect to θ.

With the derivatives of the negative log posterior we apply the HMCMC leapfrog algorithm. We can refer to the negative log posterior as E. A step size ϵ and number of iterations, L, must be defined prior to the algorithm. Thus each proposal moves a total distance of $L\epsilon$. Latent variables, p_i, are introduced for each parameter as independent normal variables with zero mean and variance M_i. Then one leapfrog step given current iteration $(\theta^b, \mathbf{p}^b)$ is

$$\begin{aligned}
\mathbf{p}^{(b+\epsilon/2)} &= \mathbf{p}^b - \frac{\epsilon}{2}\frac{dE}{d\theta}(\theta^{(b)}) \\
\theta^{(b+\epsilon)} &= \theta^b + \epsilon\frac{\mathbf{p}^{(b+\epsilon/2)}}{\mathbf{m}} \\
\mathbf{p}^{(b+\epsilon)} &= \mathbf{p}^{(b+\epsilon/2)} - \frac{\epsilon}{2}\frac{dE}{d\theta}(\theta^{(b+\epsilon)})
\end{aligned}$$

The parameters leapfrog L times ending at new proposals for the posterior. The function $H(\theta, p)$ is defined to be $E(\theta)$, which is the negative log likelihood, plus $K(p) = \frac{1}{2}\sum\frac{p_i^2}{M_i}$ The new values $(\theta^{(b+1)}, \mathbf{p}^{(b+1)})$ are accepted with probability

$$min(1, \exp\left(H(\theta^{(b+1)}, \mathbf{p}^{(b+1)}) - H(\theta^b, \mathbf{p}^b)\right)$$

. If the new value is rejected, it is set to the previous values. The method must be tuned to accept and reject at reasonable rates, perhaps accepting between 40 and 60% of proposed samples. Both L and ϵ can be tuned, where $L \times \epsilon$ is closely associated with acceptance rates.

For a normal distribution being approximated by a Hermite polynomial basis, the derivative $\frac{dB_\theta}{d\theta_i}$ is found by taking element-wise derivatives of the coefficients found in equation (4.1). The derivative is

$$\frac{db_n}{d\theta} = \frac{1}{\sigma^2} \frac{1}{\sqrt{(\sqrt{\pi} 2^n n!)\,(1+\sigma^2)}} \exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right) \sum_{k=0}^{n} H_{n,k} \left(\mu m_k - m_{k+1}\right).$$

Again, $H_{n,k}$ is the k-th coefficient in the n-th Hermite polynomial and m_k is the k-th raw moment of a normal distribution with mean $\mu/(\sigma^2+1)$ and variance $\sigma^2(\sigma^2+1)$. This can be rewritten in terms of the coefficients as

$$\frac{db_n}{d\theta} = \frac{1}{\sigma^2}\left(\mu b_n - b_{n+1} + \frac{1}{\sqrt{(\sqrt{\pi} 2^n n!)\,(1+\sigma^2)}} \exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right)\right). \tag{4.2}$$

Mixtures of normals result in more complex calculations than the normal, but the basis coefficients of the mixture as a whole is simply the sum of the individual basis coefficients. The derivative of the basis coefficients needed for the Hamiltonian MCMC is the weighted sum of the derivatives of the basis coefficients of the normal components.

4.4 Simulations

To demonstrate how the spatial Dirichlet process mixture models will work in the IDE setting with one-dimensional space, we fit the model to simulated data. Spatio-temporal data following an IDE model is simulated at 300 spatial locations and 30 time

points. The kernel is normal with mean 0 and variance 100 for locations 1 to 100, asymmetric Laplace with $\theta = 0, \mu = 5$ and $\sigma = 6$ for locations 101 to 200 and stable with $\mu = 0,\ \sigma = 3,\ \alpha = 1.3$ and $\beta = .5$ for the last 100 points. The data is shown in Figure 4.3.

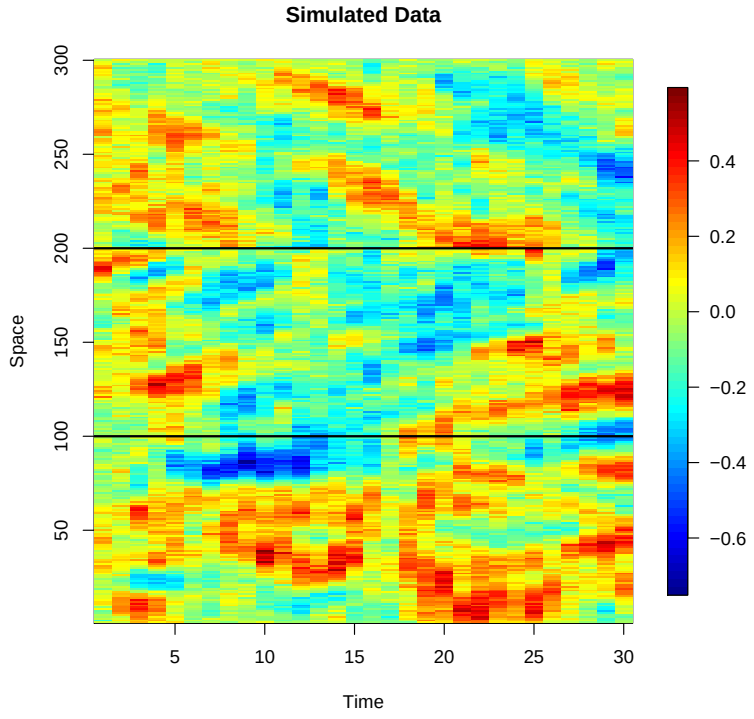


Figure 4.3: Synthetic data. The data shown is simulated using the IDE model in equations (2.8) and (2.9). The three partitioned areas use different kernels. There are 300 spatial locations and 30 time points.

The model in equations (2.8) and (2.9) is fit to the data. The parameter set for an SDP mixture kernel includes the variance of the individual component densities, σ_0^2, which will be the same value for every location and for every mixture component, the vectors of atoms $(\mu_l(s_1), ..., \mu_l(s_n))$ for $l = 1, ..., L$, and the set of latent variables $\{\xi_1, ..., \xi_{L-1}\}$ which contribute to the weights through stick-breaking. When using a Hermite function basis, the

elements of the matrix $\boldsymbol{B}_\theta$ are

$$B_\theta(n,r) = \sum_{l=1}^{L} w_l \frac{1}{\sqrt{(\sqrt{\pi}2^r r!)\,(1+\sigma_0^2)}} \exp\left(-\frac{\mu_l(s_n)^2}{2(1+\sigma_0^2)}\right) \sum_{k=0}^{r} H_{r,k} m_k^{(l)} \tag{4.3}$$

where $H_{r,k}$ is the k-th coefficient in the r-th Hermite function and $m_i^{(l)}$ is the i-th raw moment of a normal distribution with mean $\mu_l(s)/(\sigma_0^2+1)$ and variance $\sigma_0^2/(\sigma_0^2+1)$.

The number of spatially dependent parameters is reduced, using the discrete approximation of a kernel convolution (Higdon, 1998) on $\{\mu_l(s_i); i = 1, ..., n\}$, such that $\mu_l(s) \approx \int k_\zeta(u,s)\zeta_l(u)du$, where $\zeta_l(u)$ is a white noise process. Functionally, we define a kernel function and a grid, $u_1, ..., u_q$ and draw $\zeta_l(u_i) \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_\zeta^2)$, then set $\mu_l(s) = \mu_0 + \sum_{i=1}^{q} k_\zeta(u_i, s)\zeta_l(u_i)$. To ease in estimation and interpretation, a smooth Gaussian process is assumed for the kernel convolution. A Matern kernel is used with $\kappa = 2.5$ and an effective range of 40, and q is chosen to be smaller than n. The simulation uses 300 data locations, but q is set to 50. This forces smoothness which is not appropriate when modeling data, but in this application, a smooth transition of kernel shape from one location to the next is expected. This data reduction has consequences in the Hamiltonian MCMC, because inference is now needed for the latent ζ variables. Because the variables are linearly related, an application of the chain rule shows that $\partial B_\theta(n,r)/\partial\zeta_l(u_j) = \sum_{m=1}^{M} k_\zeta(u_j, s_m)\partial B_\theta(n,r)/\partial\mu_l(s_m)$

The posteriors for σ^2 and τ^2 are sampled conditionally from equations (2.11) and (2.12) using $\mathrm{IG}(3,2)$ priors. The matrix V is a Matern correlation matrix with $\kappa = 1.5$ and an effective range of 20. Then the kernel parameters σ_0^2 and $\{\xi_l : l = 1, ..., L\}$ are sampled using standard Metropolis Hastings with posterior distributions proportional to $p(a_t|a_{t-1}, \tau^2, \sigma_0^2, \{\xi_l : l = 1, ..., L\}, \{\zeta_l : l = 1, ..., L\})p(\sigma_0^2)\prod_{l=1}^{L-1} p(\xi_l)$, where the prior for σ_0^2 is a standard exponential distribution and the latent ξ_l variables have $i.i.d.$ $\mathrm{Beta}(1,\alpha)$

priors. The value for α is fixed at 2.5. Hamiltonian MCMC will be used to sample from the posterior distributions of the latent variables $\{\zeta_l(u_j); l = 1, ..., L, j = 1, ..., q\}$. There are Lq of these latent parameters, but the HMCMC is split up into L blocks, where each atom is updated individually. HMCMC is used to propose and then accept or reject $\zeta_1(u_1), ..., \zeta_1(u_q)$ as a block, then move on to $\zeta_2(u_1), ..., \zeta_2(u_q)$ and so on. The hyperparameters μ_0 and σ^2_ζ are sampled conjugately using $\mathrm{N}(0, 5)$ and $\mathrm{IG}(3, 10)$ priors respectively. The number of Hermite polynomials used is truncated to 20 and the number of atoms used is truncated to 30.

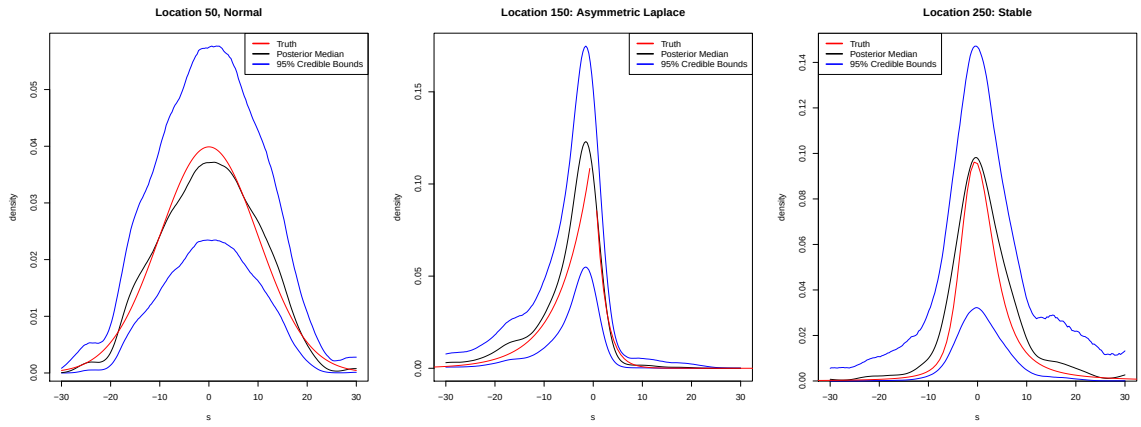


Figure 4.4: Synthetic data. Posterior mean kernel densities are shown from simulated data. The three midpoints of the regimes are chosen to display. The posterior credible interval for the kernel contains the true kernel well in each of these cases.

The posterior mean kernels for the midpoints of the three regions (locations 50, 150, and 250) are shown in Figure 4.4 along with 95% credible bands. The results show that the three different kernels are successfully recovered. The computational methods employed have allowed accurate estimation of an IDE model with a spatial DP mixture kernel. Figure 4.5 shows the advantage of using HMCMC on the DP atoms. The model has converged

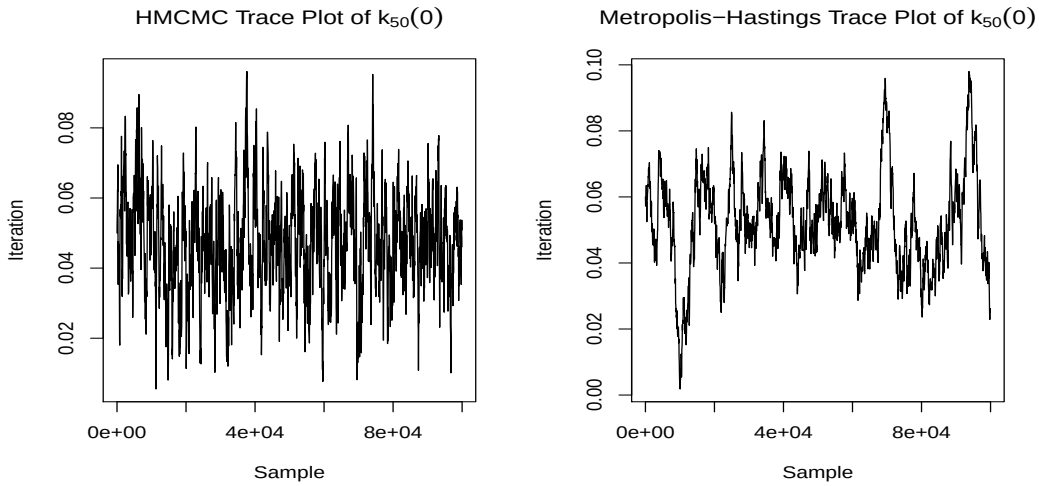


Figure 4.5: Synthetic data. Trace plots for the densities of the kernel at location $s = 50$ evaluated at $u = 0$. In the plot on the left, the atoms are updated via HMCMC and on the right they are updated using Metropolis-Hastings.

when using HMCMC at 10,000 iterations. Using Metropolis-Hastings, it is not clear if it has converged at all through 100,000 iterations, and if it has, the the chain shows a strong autocorrelation. We will now see how the model performs when applied to a real data set.

4.5 Ozone Data Analysis

To illustrate the potential of the IDE model with DP and SDP mixture kernels, we analyze the data set of ozone pressure and compare the model fits. Specifically, we will show that the SDP mixture kernel IDE model performs significantly better in prediction than any of the parametric kernels previously studied, including a spatially varying normal kernel. The data is the same as that studied using parametric kernels in Section 3.2.2. We define ozone variables for time t and location s as $Y_t(s)$ and the vector $\boldsymbol{Y}_t = \{Y_t(s_{t,1}), ..., Y_t(s_{t,n_t})\}$

where there are n_t spatial locations at time t. Similar to the model shown in section 3.2.2, the model for a general kernel with parameter set $\boldsymbol{\theta}$ is

$$\begin{aligned}
\boldsymbol{Y}_t|\boldsymbol{a}_t, \sigma^2 &\sim N(\boldsymbol{\Psi}_t\boldsymbol{a}_t + Z_{t1}^{(1)} + Z_{t2}^{(1)}, \sigma^2\boldsymbol{I}), \quad t = 1, ..., T \\
\boldsymbol{a}_t|\boldsymbol{a}_{t-1}, \tau^2, \boldsymbol{\theta} &\sim N(\boldsymbol{G}\boldsymbol{B}_\theta\boldsymbol{a}_{t-1}, \tau^2\boldsymbol{G}\boldsymbol{V}\boldsymbol{G}'), \\
\begin{pmatrix} Z_{ti}^{(1)} \\ Z_{ti}^{(2)} \end{pmatrix} &\sim N\left(\begin{pmatrix} \cos(z_i) & \sin(z_i) \\ -\sin(z_i) & \cos(z_i) \end{pmatrix}\begin{pmatrix} Z_{t-1,i}^{(1)} \\ Z_{t-1,i}^{(2)} \end{pmatrix}, \boldsymbol{W}_t^{(Z)}\right), \quad i = 1, 2 \\
\sigma^2, \tau^2, \boldsymbol{W}_t^{(Z)} &\sim p(\sigma^2)p(\tau^2)p(\boldsymbol{W}_t^{(Z)}) \\
\boldsymbol{\theta}|\gamma &\sim p(\boldsymbol{\theta}|\gamma), \qquad \gamma \sim p(\gamma).
\end{aligned}$$

The matrices $\boldsymbol{\Psi}_t$, $\boldsymbol{B}_\theta$, and $\boldsymbol{G}$ are derived from the basis function choice and the kernel choice, as described in section 2.3. The matrix V is a fixed spatial covariance matrix. To perform conditionally linear filtering for this model with the seasonal variables, we augment the state vector to $(\boldsymbol{a}_t', Z_{t1}', Z_{t_2}')'$ and augment the process level evolution matrix as a block diagonal. The model parameters σ^2, τ^2, and $\boldsymbol{W}_t^{(Z)}$ are treated similarly for each model. The parameters σ^2 and τ^2 are given IG$(3, 3)$ priors. The matrix $\boldsymbol{W}_t^{(Z)}$ is given a prior of $IW(10, 10I)$, which results in a conjugate posterior inverse Wishart distribution conditional on the state vector. The model fit using a DP mixture kernel reacts poorly to non-informative priors, but is insensitive to a wide variety of informative priors. For the filtering, priors must be defined for $\boldsymbol{m}_0$ and $\boldsymbol{C}_0$, the mean vector and covariance matrix of the time 0 state vector $\boldsymbol{a}_0$. The model may be sensitive to the specification of $\boldsymbol{m}_0$, so some care must be taken to inform a prior for the time 0 process which lies relatively near the data. For this analysis, $\boldsymbol{m}_0$ is specified to give the prior mean of the time 0 process a constant value of 3. The covariance matrix $\boldsymbol{C}_0$ is specified as $4\boldsymbol{I}$, which is considerably

diffuse for Hermite basis coefficients.

4.5.1 Dirichlet Process Mixture Kernel

This data set is analyzed using parametric kernels in section 3.2.2, where it was determined that the stable distribution performed better in prediction and scoring. The energy scores from equation (3.2) are calculated for each time point, resulting in 260 energy scores for each model. This scoring procedure will be used again to compare different models. The stable is a very flexible distribution with polynomial tails when $\alpha < 2$. The results of the previous analysis showed that, when a stable family of distributions is used for the kernel, the posterior distributions of certain parameters suggest the true kernel is skewed left. The Dirichlet process can also represent heavy tailed and skewed distributions, but it can go beyond the stable family in representing a variety of other features as well.

For the DP mixture kernel, the parameter set includes the latent variables defining the weights, $\xi_1, ..., \xi_{L-1}$, the atoms $\mu_1, ..., \mu_L$, and the kernel variance σ_0^2. The priors for the atoms is $\mathrm{N}(\mu_0, \sigma_\mu^2)$ and the prior for the latent variables are $\mathrm{Beta}(1, 2.5)$. We place hyperpriors on μ_0 of $\mathrm{N}(0, 100^2)$ and on σ_μ^2 of $\mathrm{IG}(2.5, 300)$. Again, the model is insensitive to a wide array of informative priors, but using priors which are too diffuse can delay convergence of the HMCMC. Figure 4.6 shows the estimated posterior means of the densities.

The two densities are very similar. The main difference is the thickness of the left tail. The DP mixture kernel model scored lower than the stable distribution model 64% of all time points, suggesting that there is an advantage to using the DP mixture, but perhaps for many situations, it may not be enough of an advantage considering the extra parameters. Of course, if this data is non-stationary, then these results will be greatly

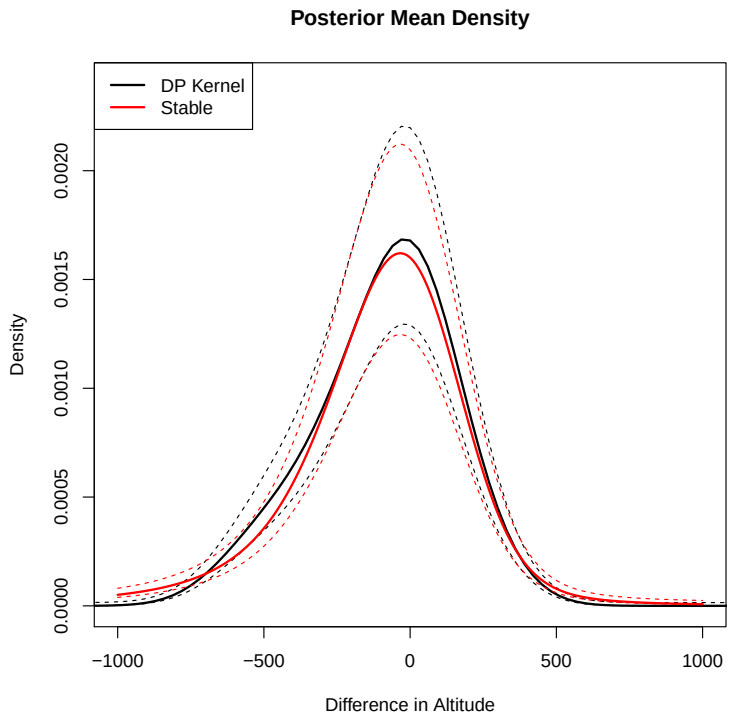


Figure 4.6: Ozone data. The posterior mean densities are shown for the stable distribution and the Dirichlet process mixture of normals kernels, with 95% credible bands. The two fits are very close to each other.

affected. Further studies on stationary data sets will reveal more information on how these two models compare. To improve upon the model fit for ozone data, the SDP mixture kernel should be used.

4.5.2 Spatial Dirichlet Process Mixture Kernel

In all, six different kernels are used to fit the IDE model to the ozone pressure data: normal, asymmetric Laplace, stable, Dirichlet process mixture of normals, spatially varying normal, and spatial Dirichlet process mixture of normals. The size of the kernel parameter set, $\boldsymbol{\theta}$, for this analysis varies from 2 for the normal to over 1,000 for the SDP mixture. The spatially varying normal kernel IDE model uses a convolution described in the section 4.4 simulation for both the mean and the log variance. Again q is set to 50, which is to say there are 50 locations to place knots $\boldsymbol{\zeta} = (\zeta(u_1), ..., \zeta(u_q))'$, which are distributed as normal random variables with mean 0 and variance σ^2_ζ. Then a discretized version of the kernel convolution is used by assigning the mean process at points $s_1, ..., s_n$ to $(\mu(s_1), ..., \mu(s_n))' = \boldsymbol{K\zeta} + \mu_0\mathbf{1}$. The random variables in $\boldsymbol{\zeta}$ will be estimated using HMCMC. By construction, the prior for each ζ random variable is $\mathrm{N}(0, \sigma^2_\zeta)$. A similar treatment is used for the log variance, with latent variables $\boldsymbol{\eta} = (\eta(u_1), ..., \eta(u_q))'$ which are $i.i.d.$ from $\mathrm{N}(0, \sigma^2_\eta)$. The log variance is set to $(\log(\sigma^2(s_1)), ..., \log(\sigma^2(s_n)))' = \boldsymbol{K\eta} + \mu_\sigma\mathbf{1}$.

For the spatially varying Gaussian kernel and the SDP mixture kernel, the prior mean μ_0 has a $\mathrm{N}(0, 50^2)$ hyerprior and the variance of the SDP mixture kernel, σ^2_0, has a $\mathrm{Gamma}(10, .1)$ prior. An $\mathrm{IG}(3, 100)$ prior is placed on σ^2_ζ and σ^2_η, and a $\mathrm{N}(100, 100)$ prior is used for μ_σ. While extensions can be made, for simplicity in the illustration α is fixed at 2.5 and the DP is truncated to 30 atoms. The location of the knots used in the convolution and

the amount chosen may significantly impact the model. Choosing too few knots or failing to place knots outside the spatial boundary of the data can affect how well the model performs in certain regions.

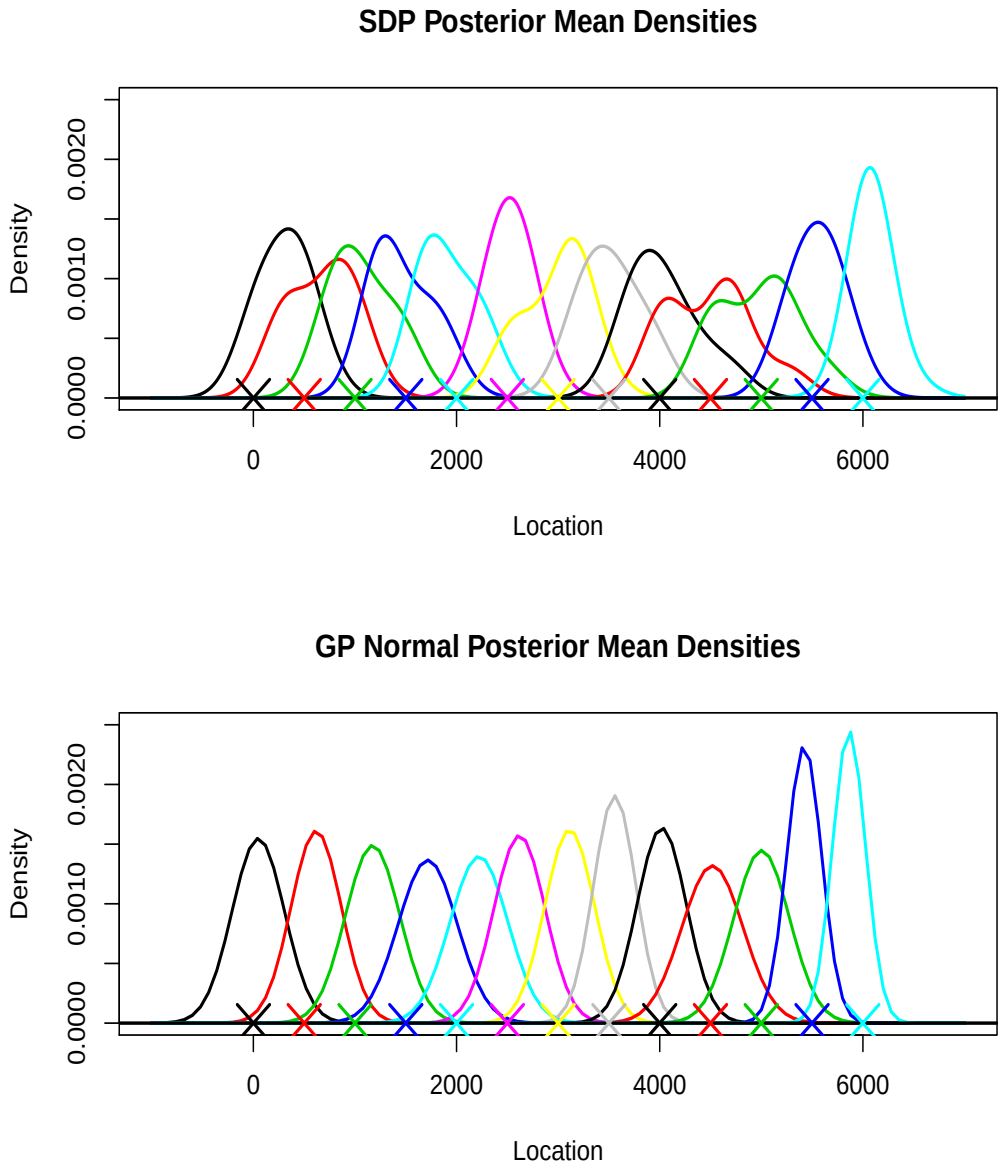


Figure 4.7: Ozone data. The curves are estimated posterior means of the kernel densities for the spatial DP mixture kernel IDE. The X's on the x-axis show the spatial location associated with the kernel of the matching color.

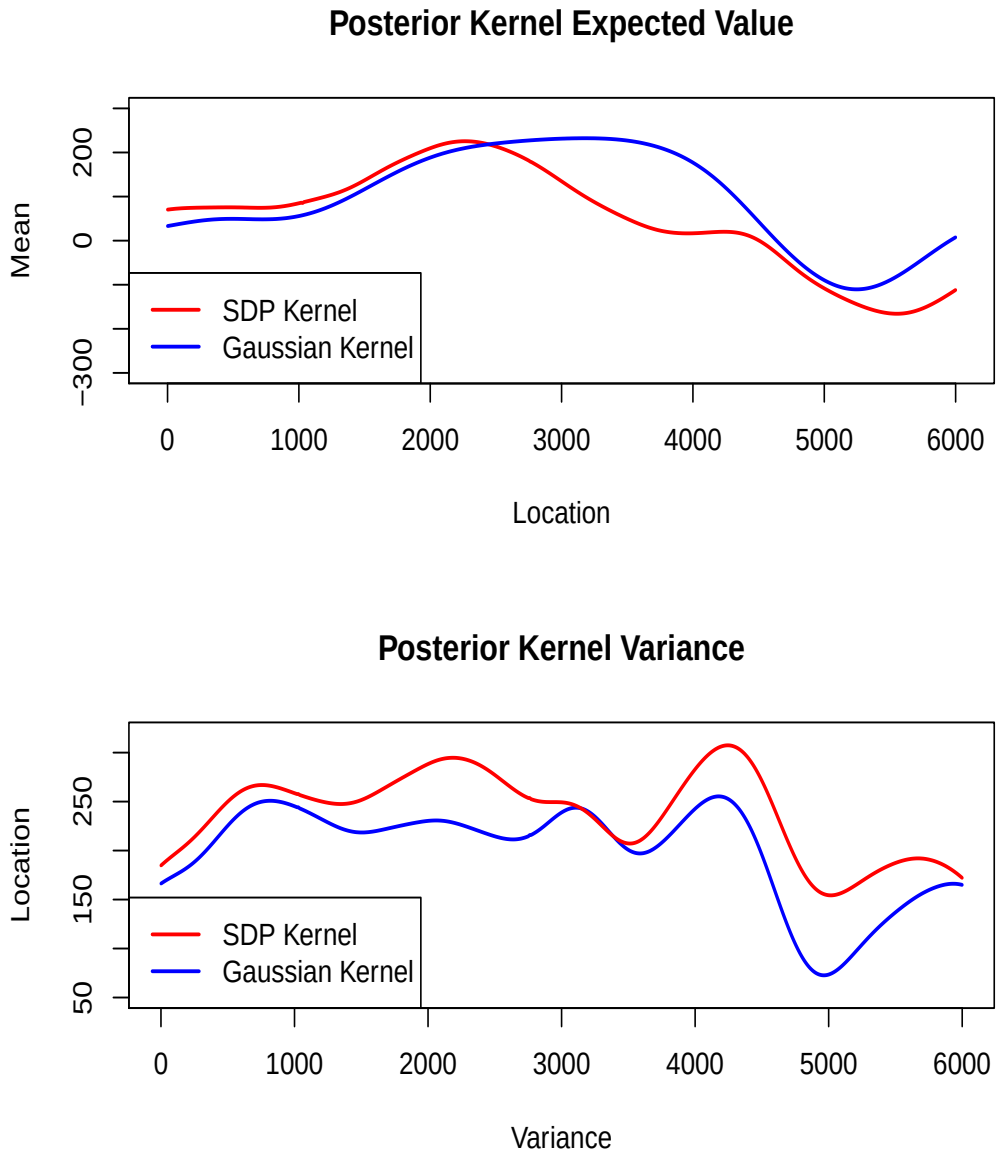


Figure 4.8: Ozone data. Mean expected value and variance of the sampled SDP mixture kernel and spatially varying normal kernel across the locations of the data.

20,000 samples from the posterior were taken after careful tuning of the Hamiltonian MCMC and Metropolis-Hastings steps. The estimated posterior mean of the kernel

density for the SDP mixture model will change across space. Figure 4.7 shows the posterior mean point estimates of the kernel for an array of spatial locations. The X's on the x-axis show the spatial location and the density with the matching color is the estimated posterior mean kernel at that location. The kernel shifts skewness from left to right throughout the range of the data. The kernels in the higher altitudes of the data have heavier tails than the kernels in the lower altitudes. Also, some bimodalilty is seen. While bimodality may be tough to interpret in a physical sense, it is a feature which would be impossible to recreate using a less flexible kernel. To compare with the spatially varying normal kernel, Figure 4.8 shows how the expected value and variance of the sampled kernels vary across space. There is a clear association. The variance of the SDP mixture, however, is consistently larger than the spatially varying normal. One-step ahead prediction profiles for 3 different time points are shown for all 6 of the models in Figure 4.9. We see how the model improves with the flexibility of the kernel.

For each time point we calculate the energy scores from equation (3.2) and compare. Lower is better for these energy scores and for 222 of the 260 time points, the spatial DP mixture of normals kernel IDE model has the lowest energy score. The spatially varying normal kernel IDE model has the lowest score for 20 of the remaining time points. The stable and the Dirichlet process kernels scored lowest 9 times each and the stationary normal and asymmetric Laplace never did. The stable and DP mixture models did have lower scores than the Gaussian process mean kernel 12% of the time. From the profiles and the scoring procedures, it is clear that using spatially varying parameters is advantageous despite the difficulty of learning the complicated models. Also, the spatial DP mixture model

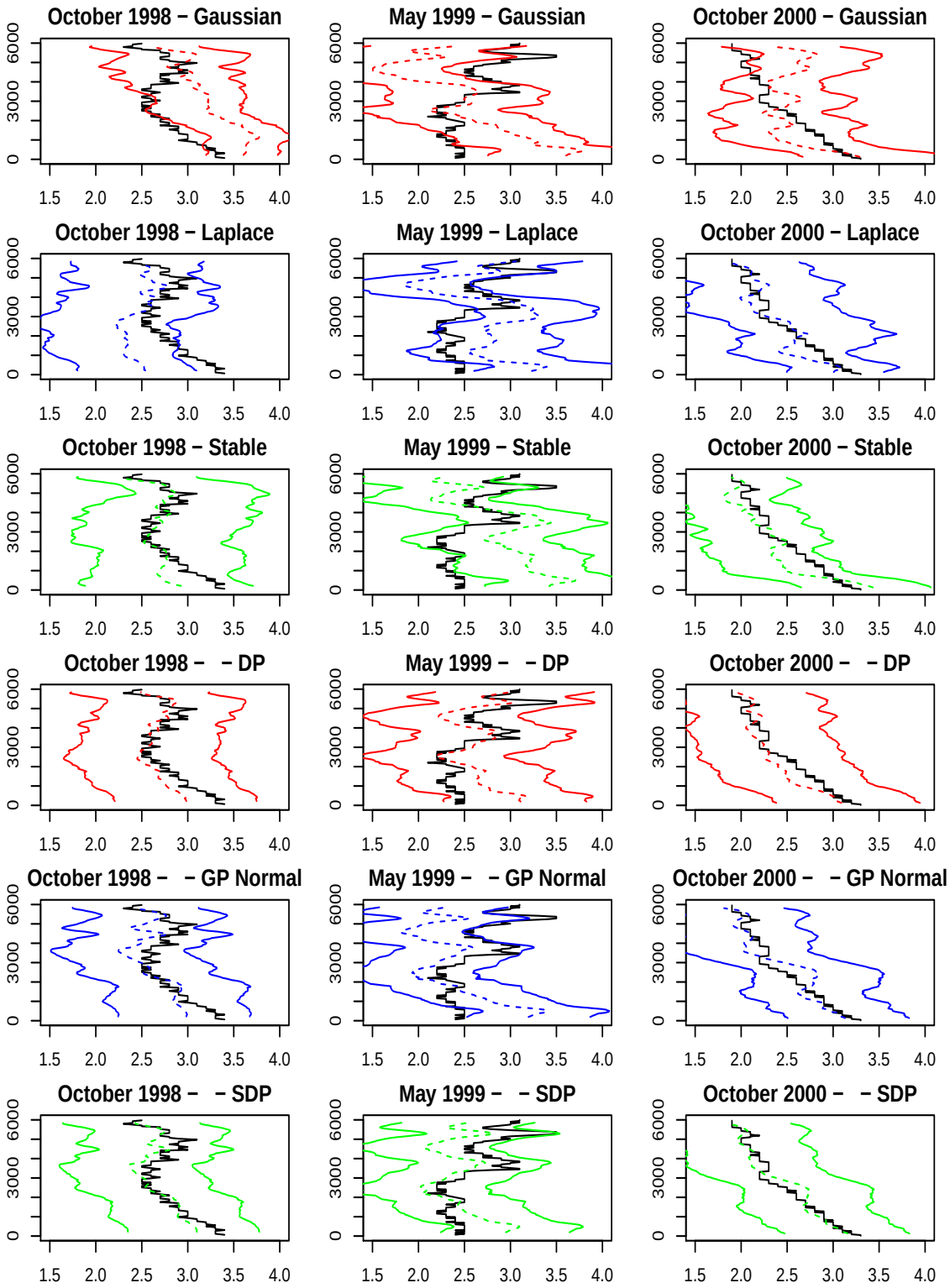


Figure 4.9: Ozone data. Profiles for one-step ahead predictions for all 6 models are shown for 3 time points.

clearly performs the best. Figure 4.10 shows the fitted values and residuals for the SDP mixture model.

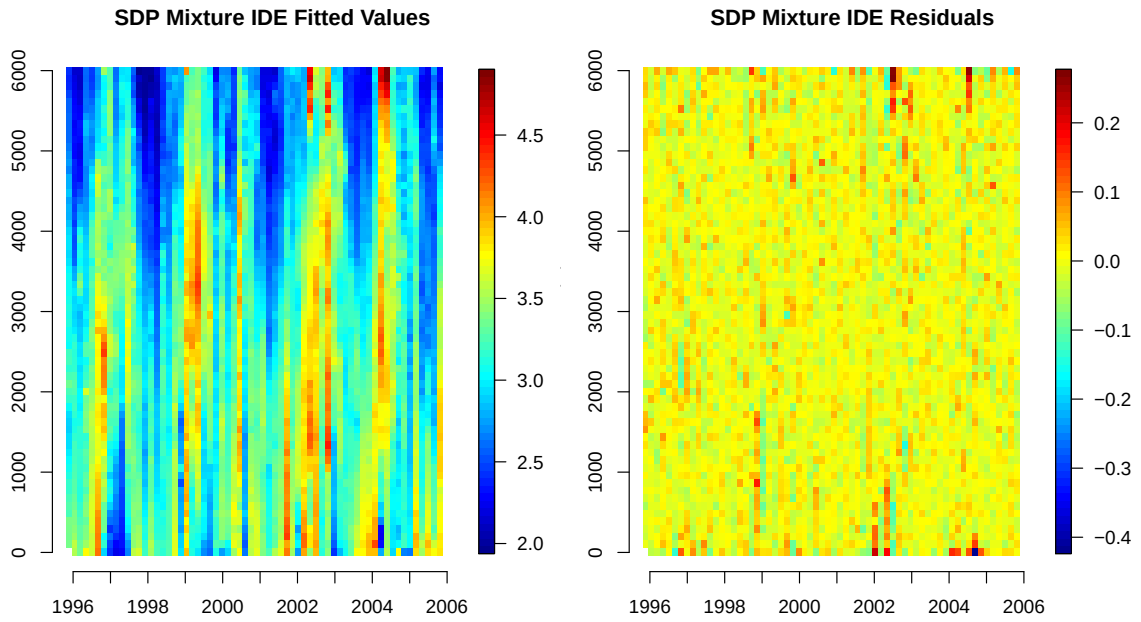


Figure 4.10: Ozone data. On the left are the fitted values for the SDP mixture kernel IDE model. The right plot shows residuals for the IDE model.

The frequencies of the harmonics were chosen by comparing model fits when using the parametric kernels, but the non-stationary models or the DP mixture kernel models may require different harmonics or more of them. By using harmonics the resulting forecast function includes a cyclical sinusoidal element. The amplitude and phase of this forecast function for the SDP kernel IDE model can show how the harmonics affect the model. According to West and Harrison (1997), we find the amplitude from the state variables as $\sum_{i=1}^{2} \sqrt{\sum_{j=1}^{2} Z_{ti}^{(j)2}}$. The amplitude of the first harmonic averages 0.108 for all time points and decreases slightly from 1996 to 2006. The second harmonic averages 0.76 and increases

slightly over the time span. The values for the phase shift are $\arctan(-Z_{ti}^{(2)}/Z_{ti}^{(1)})$. The posterior means for the phase vary randomly about 0.

4.6 Conclusion

We have explored the full potential of IDE models by using Bayesian nonparametric kernels. Despite the computational concerns, when the model can be fit properly, it is very powerful. The spatial Dirichlet process mixture of normals kernel is able to capture a variety of spatio-temporal effects which are unable to be recovered using other kernels. For our ozone example, we saw our model switch the direction it was skewed several times. There may be some underlying physical explanation as to why we observe this. Also, different regions can have heavier tails. Capturing complicated tail behavior may be easier when it is allowed to change over space instead of coercing the tails of all locations to be equal. Scoring procedures and profile plots have shown that for prediction, the SDP mixture kernel IDE model performed better than the stationary IDE models. The spatially varying normal kernel also seemed to perform better than the stationary kernel IDE models, but not as convincingly. To avoid the many potential pitfalls to fitting these models, we propose careful consideration of basis function, estimation technique, as well as truncations points. The Hermite basis function appears to be a good basis choice for a mixture of normals, if used properly. We have proposed Hamiltonian MCMC updates for the atoms of the Dirichlet process mixture.

Extending this to two dimensions has proven to be difficult. Any computational problems in one-dimensional space are much more severe in two-dimensional space. Simu-

lations suggest that not even HMCMC techniques are able to learn the SDP mixture kernel parameters. Some other advanced sampling methods may be able to learn the parameters in this case.

Chapter 5

Bivariate Stable Kernel IDE Modeling

5.1 Introduction

Previous chapters have argued the value of using flexible kernels in IDE modeling and have provided strong evidence that it results in improvement of model accuracy and prediction. To help in interpretation of the more complex kernels and to provide a more complete illustration of the added benefits, these arguments were made in one-dimensional space. There are a number of practical concerns for extending the model to two dimensions. Convergence of parameters has proven difficult for non-parametric 2-dimensional kernels, and reproducing kernels used for simulated data has not been achieved. While one-dimensional data sets exist, such as the ozone data presented in chapters 3 and 4, river data, and ocean depths, the majority of spatio-temporal data sets are measured in

two-dimensional space. In this chapter we propose flexible kernel IDE modeling for two dimensions. We achieve this by using a stable kernel, which is defined through a measure, Γ, which controls characteristics of the shape of the kernel. Flexible modeling of the measure results in a large variety of kernel shapes. While estimation of stable kernel IDE models is more feasible than non-parametric kernels, it is not without challenges. Flexibly modeling the measure itself must be carefully considered.

Section 5.2 covers some relevant details which will be used in the fitting of IDE models with stable distributions. Methods of modeling the measure and learning the model are given in section 5.3. Section 5.4 details an analysis of sea surface temperature data with the focus on prediction. We show that the IDE model with a stable kernel improves prediction over the normal kernel IDE model.

5.2 Theory

Several details must be specified for successful IDE modeling using bivariate stable kernels. We begin with definitions in section 5.2.1, including the connection between the measure, Γ, and the shape of the kernel. To measure the advantages of using a stable kernel over the elliptically symmetric Gaussian kernels, symmetry considerations of the stable kernel are discussed in section 5.2.2. Finally, details of using real-valued Fourier series for the bivariate stable density are detailed in section 5.2.3.

5.2.1 Stable Distributions

A bivariate vector $\boldsymbol{X}$ belongs to the stable family of distributions if and only if for any constants A and B and independent copies $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, there exists a constant C and vector $\boldsymbol{D} \in \mathbb{R}_2$ such that $A\boldsymbol{X}^{(1)} + B\boldsymbol{X}^{(2)} = C\boldsymbol{X} + \boldsymbol{D}$ (Samorodnitsky and Taqqu, 1997). Many additional properties stem from this definition, such as infinite divisibility and the lack of finite moments. As a family of distributions with no finite moments, the stable family has been used for a variety of infinite variance applications, such as financial data (Nolan, 2014; Panorska, 1996) and signal processing with heavy-tailed noise (Nolan et al., 2010). The one-dimensional stable distribution is described in section 3.1.2. In one dimension, the stable family is defined by 4 parameters: $\mu \in \mathbb{R}$ is a location parameter, $c > 0$ is a scale parameter, $\alpha \in (0, 2]$ controls the thickness of the tails, and $\beta \in [-1, 1]$ controls skewness. The characteristic function in one dimension is given in equation (3.1). The multivariate stable characteristic function for $\alpha \neq 1$ is

$$g(\boldsymbol{t}|\alpha, \boldsymbol{\mu}, \Gamma) = \exp\left\{ i\boldsymbol{t}'\boldsymbol{\mu} + \int_{\boldsymbol{s}\in S_d} |\boldsymbol{t}'\boldsymbol{s}|^{\alpha}(1 - i\ \text{sign}(\boldsymbol{t}'\boldsymbol{s})\tan(\pi\alpha/2))\Gamma(d\boldsymbol{s}) \right\}. \tag{5.1}$$

Relating this to the one-dimensional stable distribution, the vector $\boldsymbol{\mu}$ is a location parameter, α controls tail behavior, and the measure Γ controls characteristics of the distribution such as skewness, orientation, and spread, effectively replacing both c and β. For a stable vector of size d, the integration space S_d is the unit sphere is $\mathbb{R}_d$. For 2 dimensions, S_2 is the unit circle. For this special case, a change of variables can be made to $\boldsymbol{s} = (\cos(z), \sin(z))'$ where the integral will now be taken over $z \in [0, 2\pi]$. The unscaled density of the measure, $\gamma = d\Gamma$, must be non-negative, but can take on a variety of other shapes. Figure 5.1 shows how the shape of the distribution changes with γ. The skewness of the data set changes

when $\gamma(z)$ and $\gamma(z+\pi)$ are more disparate. The spread of the distribution changes with the scale of the measure, meaning a measure which is larger for all z will have a larger spread. The orientation of the distribution will rotate with a shift in γ. A variety of other distributional shapes can be achieved by combining these properties.

5.2.2 Symmetry

Elliptically contoured stable distributions are a simplification of stable laws (Nolan, 2013) and have characteristic functions of the form $\exp\left(-(\boldsymbol{t}'\Sigma\boldsymbol{t})^{\alpha/2} + i\boldsymbol{t}'\boldsymbol{\mu}\right)$. Tasks which have proven difficult for the multivariate stable, such as evaluating the density and estimating parameters, are simplified when using elliptically contoured stable distributions. These are also referred to as sub-Gaussian distributions because they can be represented as a scale mixture of normals. A stable distribution is called symmetric α-stable if there exists a $\boldsymbol{\mu}$ such that $-\boldsymbol{X} + \boldsymbol{\mu} \stackrel{d}{=} \boldsymbol{X} + \boldsymbol{\mu}$. This property is equivalent to that of elliptical symmetry, although elliptical symmetry is often defined using the first and second moments, which the stable distribution will not have. A major advantage the stable kernel IDE model will have over the Gaussian kernel IDE model is that it can achieve skewness, thus it will be important to measure how skewed a given stable distribution is. There is a specific form of Γ which ensures elliptically contoured stable laws (see equation 2.5.8 in Samorodnitsky and Taqqu (1997)). The measure Γ can also determine elliptical symmetry.

Lemma 6. *Let $\boldsymbol{X}$ follow a bivariate stable distribution with location parameter $\boldsymbol{\mu}$ and measure Γ. The property that $\gamma(z) = \gamma(z+\pi)$ for $z \in [0, \pi]$ is necessary and sufficient for the distribution of $\boldsymbol{X}$ to be elliptically symmetric.*

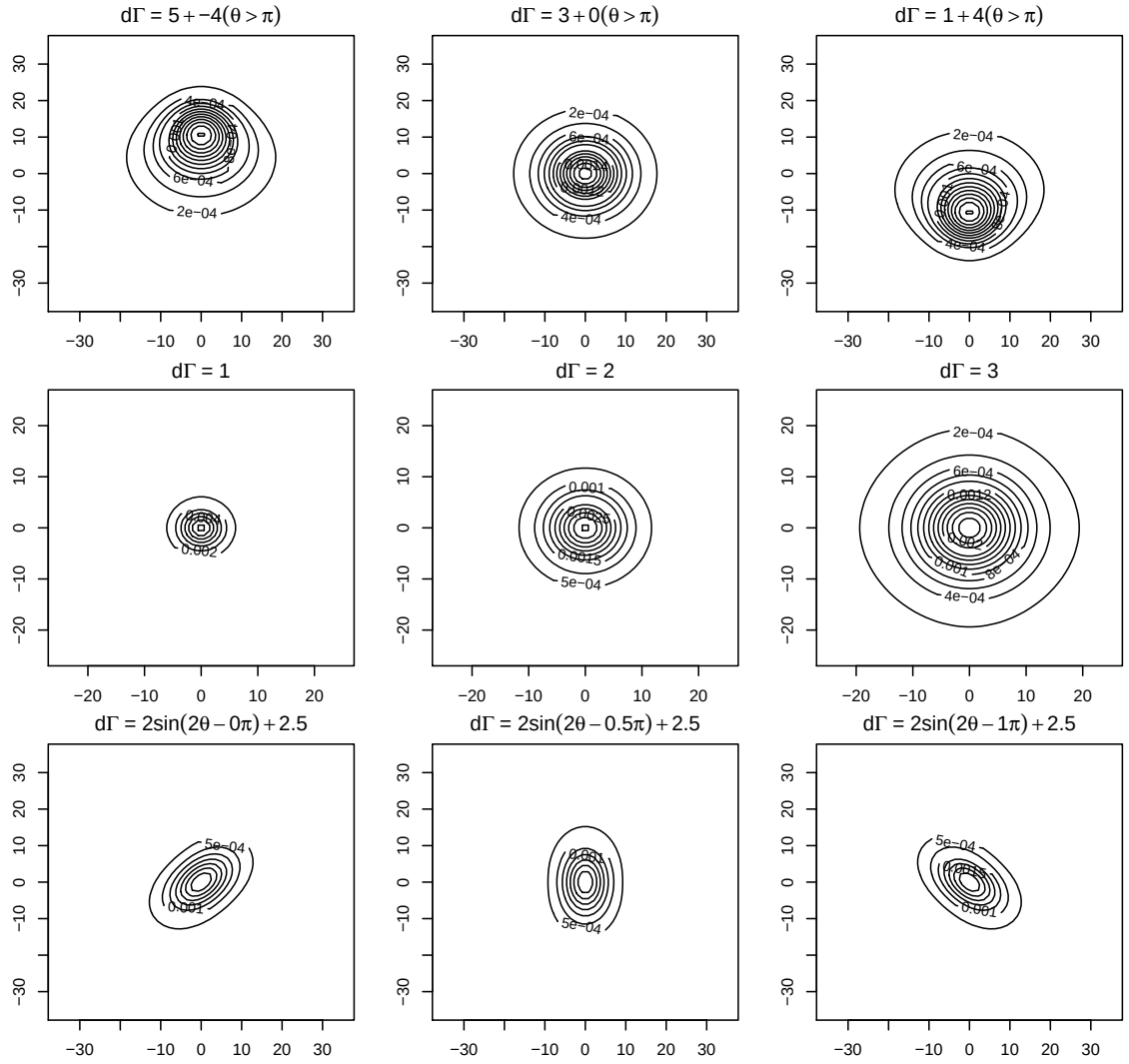


Figure 5.1: The shape of the bivariate stable changes with Γ. For all these plots, $\boldsymbol{\mu} = (0, 0)'$ and $\alpha = 1.5$. In the top row γ is a changing step function resulting in a skewed distribution. In the middle row, γ is a changing constant function resulting in different spreads of the distribution. In the bottom row γ is a sine function with a changing shift, resulting in different orientations of the distribution.

Proof. Without loss of generality, assume $\boldsymbol{\mu} = 0$. Using a Fourier representation of the density, $f(\boldsymbol{X}) = (2\pi)^{-2} \sum_{\boldsymbol{t} \in \mathbb{Z}_2} g(t|\alpha, \Gamma) e^{i\boldsymbol{t}'\boldsymbol{X}}$. The value $f(\boldsymbol{X}) - f(-\boldsymbol{X})$ can be simplified to $(2\pi)^{-2} \sum_{\boldsymbol{t} \in \mathbb{Z}_2} \left(g(t|\alpha, \Gamma) - g(-t|\alpha, \Gamma)\right) e^{i\boldsymbol{t}'\boldsymbol{X}}$ using a simple transformation. First assume that the distribution is symmetric, which implies $f(\boldsymbol{X}) - f(-\boldsymbol{X}) = 0$. Properties of Fourier and inverse Fourier transforms show that this is equivalent to $g(t|\alpha, \Gamma) - g(-t|\alpha, \Gamma) = 0$ for all $t \in \mathbb{Z}_2$. By collecting terms and dividing constants, this becomes $\int_{z=0}^{2\pi} \text{sign}(\boldsymbol{t}'\boldsymbol{s}) |\boldsymbol{t}'\boldsymbol{s}|^{\alpha} \Gamma(dz) = 0$. The integral can be split and simplified into $\int_{z=0}^{\pi} |\boldsymbol{t}'\boldsymbol{s}|^{\alpha} d\Gamma(z) - \int_{\theta=0}^{\pi} |\boldsymbol{t}'\boldsymbol{s}|^{\alpha} d\Gamma(z + \pi) = 0$ for all $t \in \mathbb{Z}_2$, which holds true only when $\gamma(z) = \gamma(z + \pi)$ for $z \in [0, \pi]$.

To find the reverse, assume that $\gamma(z) = \gamma(z + \pi)$. By working backwards in the above proof, this implies $\int_{z=0}^{2\pi} \text{sign}(\boldsymbol{t}'\boldsymbol{s}) |\boldsymbol{t}'\boldsymbol{s}|^{\alpha} \Gamma(dz) = 0$. The characteristic function then becomes $g(\boldsymbol{t}|\alpha, \boldsymbol{\mu}, \Gamma) = \exp\left\{ \int_{z=0}^{2\pi} |\boldsymbol{t}'\boldsymbol{s}|^{\alpha} \tan(\pi\alpha/2)) \Gamma(dz) \right\}$. Then $g(t|\alpha, \Gamma) - g(-t|\alpha, \Gamma) = 0$ and $f(\boldsymbol{X}) - f(-\boldsymbol{X}) = 0$, resulting in the property of elliptical symmetry. □

Varying degrees of skewness can be achieved with more or less disparate values of $\gamma(z)$ and $\gamma(z + \pi)$.

5.2.3 2-Dimensional Fourier Series

The typical method for fitting an IDE model to spatio-temporal data involves decomposing the process and the kernel into an orthonormal basis series expansion. A frequent choice is the Fourier series, due to the connection between the Fourier transform for probability densities and the characteristic function. In one dimension, the Fourier basis functions $\{\exp(ikx) : k = 0, \pm 1, \pm 2, ...\}$ are often replaced with real valued functions $\cos(kx)$ and $\sin(kx)$ for $k = 0, 1, 2, ...$ The replacement of the bivariate Fourier complex-

valued basis functions with real-valued functions is not used as frequently in literature. If the characteristic function of a density is $g(\boldsymbol{t})$ then the function can be represented as $f(\boldsymbol{x}) = (2\pi)^{-2} \sum_{t_1=0,\pm1,\ldots} \sum_{t_2=0,\pm1,\ldots} e^{i\boldsymbol{t}'\boldsymbol{x}} g(\boldsymbol{t})$. When $g(\boldsymbol{t}) = \exp(a(\boldsymbol{t}) + ib(\boldsymbol{t}))$ the real basis functions are $\cos(\boldsymbol{t}'\boldsymbol{x})$ and $\sin(\boldsymbol{t}'\boldsymbol{x})$ and the attached coefficients of the expansion are respectively $\exp(a(\boldsymbol{t}))\cos(b(\boldsymbol{t}))$ and $\exp(a(\boldsymbol{t}))\sin(b(\boldsymbol{t}))$. The basis coefficients can be simplified further by combining the positive and negative indices of the expansion. The basis functions would remain the same, but only the indices where $t_1 \geq 1$ for all t_2 or for $t_2 \geq 0$ when $t_1 = 0$ will be included in the basis function set. The new coefficients of these basis functions in an expansion for a density will be $\exp(a(\boldsymbol{t}))\cos(b(\boldsymbol{t})) + \exp(a(-\boldsymbol{t}))\cos(b(-\boldsymbol{t}))$ and $\exp(a(\boldsymbol{t}))\sin(b(\boldsymbol{t})) - \exp(a(-\boldsymbol{t}))\sin(b(-\boldsymbol{t}))$ except for when $j = k = 0$, for which case the basis coefficient is $\exp(a(\boldsymbol{0}))\cos(b(\boldsymbol{0}))$. The specific application of the Fourier series expansions used for IDE modeling requires this simplification to make the matrix which maps the basis coefficients to the data to be full rank.

The stable distribution is used in a variety of applications where the density is required. This has led to a number of efforts to approximate or recreate the kernel in some way. These include discretization of the Γ measure (Byczkowski et al., 1993), and using one-dimensional projections (Abdul-Hamid and Nolan, 1998; Matsui and Takemura, 2009). A motivation for these approaches is computational feasibility in higher dimensions. The lack of scalability of Fourier series approximations could be a reason why it is not commonly used to represent the stable density. For the specific application of the IDE, only a bivariate stable is required, so a Fourier series approximation will be a reasonable method of approximating the density. The characteristic function for the multivariate stable in equation (5.1) can be

written for the bivariate case as $\exp\left(a(\boldsymbol{t}) + ib(\boldsymbol{t})\right)$ where $a(\boldsymbol{t}) = \int_{z=0}^{2\pi} |\boldsymbol{t}'\boldsymbol{s}|^{\alpha} \tan\left(\pi\alpha/2\right)\Gamma(dz)$ and $b(\boldsymbol{t}) = \boldsymbol{t}'\boldsymbol{\mu} - \int_{z=0}^{2\pi} \text{sign}(\boldsymbol{t}'\boldsymbol{s})|\boldsymbol{t}'\boldsymbol{s}|^{\alpha} \tan\left(\pi\alpha/2\right)\Gamma(dz)$. Recall that $\boldsymbol{s} = (\cos(z), \sin(z))'$. Table 1 reviews the properly scaled Fourier basis functions and coefficients for the bivariate stable distribution. These basis coefficients are for $t_1 \geq 1$ for all t_2 and for $t_2 \geq 1$ when $t_1 = 0$. The basis coefficient when $t_1 = t_2 = 0$ is $(2\pi)^{-1}$, and only the cosine basis function should be included.

Basis Function	Coefficient
$(2\pi)^{-1}\cos\left(\boldsymbol{t}'\boldsymbol{x}\right)$	$(\pi)^{-1}\exp\left(\int_{z=0}^{2\pi} \lvert\boldsymbol{t}'\boldsymbol{s}\rvert^{\alpha} \tan\left(\pi\alpha/2\right)\Gamma(dz)\right)\cos\left(\boldsymbol{t}'\boldsymbol{\mu} - \int_{z=0}^{2\pi} \text{sign}(\boldsymbol{t}'\boldsymbol{s})\lvert\boldsymbol{t}'\boldsymbol{s}\rvert^{\alpha} \tan\left(\pi\alpha/2\right)\Gamma(dz)\right)$
$(2\pi)^{-1}\sin\left(\boldsymbol{t}'\boldsymbol{x}\right)$	$(\pi)^{-1}\exp\left(\int_{z=0}^{2\pi} \lvert\boldsymbol{t}'\boldsymbol{s}\rvert^{\alpha} \tan\left(\pi\alpha/2\right)\Gamma(dz)\right)\sin\left(\boldsymbol{t}'\boldsymbol{\mu} - \int_{z=0}^{2\pi} \text{sign}(\boldsymbol{t}'\boldsymbol{s})\lvert\boldsymbol{t}'\boldsymbol{s}\rvert^{\alpha} \tan\left(\pi\alpha/2\right)\Gamma(dz)\right)$

Table 5.1: Basis functions and coefficients for a real-valued Fourier basis expansion of the bivariate stable distribution.

5.3 Methods of Posterior Inference

Using the basis function decomposition of the process and the distribution, the IDE model with a stable kernel can be represented by equations (2.8) - (2.10). For the bivariate stable kernel, the parameter set, $\boldsymbol{\theta}$, includes $\boldsymbol{\mu}, \alpha$, and Γ. A rich body of literature has been dedicated to learning the parameters of the stable distribution. These methods are typically data-based, such as forming an empirical characteristic function (Nolan et al., 2001), or empirical likelihoods (Ogata, 2013). A Bayesian treatment of estimating stable parameters is found in Salas-Gonzalez et al. (2010), though it is restricted to modeling mixtures of symmetric α-stable distributions. Estimation can be simplified by restricting to

elliptically contoured stables (Nolan, 2013). Data based methods will not be possible with how the distribution is embedded into the IDE model. Using the Fourier representation, likelihood based inference methods are possible for the kernel parameters.

5.3.1 Bernstein Polynomials

We place a Bernstein polynomial basis prior on Γ to produce a flexible construction (Petrone, 1999). Because Γ need not be a proper probability distribution, a scale parameter which contributes to the spread of the distribution is added. The measure can be written as

$$\gamma(\theta) = c \sum_{m=1}^{M} w_{M,m} B(\theta; m, M - m + 1)$$

where $B(\theta/2\pi; m, M - m + 1)$ is a Beta density function. The weights are deterministic given a base distribution F. They are

$$w_{M,m} = F\left(\frac{m}{M}\right) - F\left(\frac{m-1}{M}\right).$$

By making F random and choosing a large value of M, the resulting measure becomes a non-parametric prior on the space of 0 to 2π. For example, the weights can be realizations from a Dirichlet distribution with parameters $\{\alpha\left(F_0\left(\frac{m}{M}\right) - F_0\left(\frac{m-1}{M}\right)\right), m = 1, ..., M\}$, where F_0 is some base distribution. With this definition of γ, a very large set of probability measures can be used.

Because these parameters are learned in an MCMC setting while embedded into a complicated model, learning the weights may be difficult with a Dirichlet process motivating the base distribution. To simplify the model while maintaining a great deal of flexibility, we use a geometric weights prior for F (Mena et al., 2011). This is done by drawing atoms

$x_1, ..., x_J$ from a uniform on 0 to 2π and assigning weight $q(1-q)^{j-1}$ to atom x_j, where $q \sim \text{Beta}(1, \alpha)$. Combining equations (2.8) and (2.9) with details about the kernel the full IDE model is

$$\begin{aligned}
\boldsymbol{Y}_t|\boldsymbol{a}_t, \sigma^2 &\sim N(\boldsymbol{\Psi}_t\boldsymbol{a}_t, \sigma^2\boldsymbol{I}_t), \quad t = 1, ..., T && (5.2)\\
\boldsymbol{a}_t|\boldsymbol{a}_{t-1}, \tau^2, \boldsymbol{\theta} &\sim N(\boldsymbol{G}_t\boldsymbol{B}_{\theta,t}\boldsymbol{a}_{t-1}, \tau^2\boldsymbol{G}_t\boldsymbol{V}_t\boldsymbol{G}_t') && (5.3)\\
\gamma(z) &= c\sum_{k=1}^{K} w_{M,m}\text{Beta}(z/2\pi|m, M-m+1)/2\pi && (5.4)\\
w_{M,m} &= F\left(\frac{m}{M}\right) - F\left(\frac{m-1}{M}\right), \quad f(\cdot) = \sum_{j=1}^{J} q(1-q)^{j-1}\delta_{x_j}(\cdot) && (5.5)\\
q &\sim \text{Beta}(1, \alpha), \quad x_j \overset{i.i.d.}{\sim} Un(0, 2\pi), \quad \boldsymbol{\mu}, c \sim p(\boldsymbol{\mu})p(c). && (5.6)
\end{aligned}$$

The latent x_j variables only enter into the model by assigning weight $q(1-q)^{j-1}$ to whichever region, $((m-1)/M, m/M)$, it lies in. Thus we can reparameterize the Bernstein polynomial weights as

$$\begin{aligned}
w_{M,m} &= \sum_{j=1}^{J} q(1-q)^{j-1}Z_{jm} && (5.7)\\
(Z_{j1}, ..., Z_{jM}) &\sim \text{Multinomial}(1, (1/M, ..., 1/M)), \quad j = 1, ..., J && (5.8)
\end{aligned}$$

This replaces a continuous latent variable with a discrete variable, which will aid in the estimation. By construction, only one of $Z_{j,1}, ..., Z_{j,M}$ will be 1 and the rest will be 0, so the dimensionality of the new parameter set is effectively the same as it was before. The parameters in Γ are c, q, and $\{Z_{j,m}, j = 1, ..., J, m = 1, ..., M\}$.

To construct a non-stationary spatio-temporal process, the parameters $\boldsymbol{\mu}$, c, and q will be spatially varying. The result will be Gaussian process priors on $\boldsymbol{\mu}(\boldsymbol{s})$ and $\log(c(\boldsymbol{s}))$. To generate a spatially varying geometric weight, we remove the Beta prior and make a

latent process $u(s)$ which has a Gaussian process prior. Then $q(s) = \phi(u(s))$, where ϕ is the standard normal distribution function. The latent assignment variables $\{Z_{j,m}, j = 1, ..., J, m = 1, ..., M\}$ will not be spatially varying. Figure 5.2 shows how the kernel can change with q. Even though the atoms are the same for each kernel, the shape can drastically change in both skewness and orientation by only varying q. Also, Figure 5.2 shows that the kernel will vary smoothly with q, so because kernels at nearby locations will have more similar values for q, the kernel shape will be more similar. This smooth transition of kernel shape is what will be expected for spatio-temporal models. One computational advantage this model will have compared to the spatially varying Gaussian kernel IDE model is that it can achieve a wider array of kernel shapes with one fewer spatially varying parameter.

5.3.2 Posterior Sampling

The state vectors $\{a_0, ..., a_T\}$ will be sampled via dynamic linear model filtering as shown in section 2.4. The kernel parameters will be updated using MCMC sampling. In the spatially varying case, the spatial process parameters will be calculated via a kernel convolution (Higdon, 1998) for μ_1, μ_2, c, and q. There will be a grid of knots $u_1, ..., u_Q$ and latent variables $\boldsymbol{\zeta}_{\mu_1} = \zeta_{\mu_1}(u_1), ..., \zeta_{\mu_1}(u_Q)$ which are *i.i.d.* $N(0, \sigma_{\mu_1})$. Then the process will be $\mu_1(s) = \mu_{\mu_1} + \sum_{i=1}^{Q} k_\zeta(u_i, s)\zeta_{\mu_1}(u_i)$. Using Gibb's sampling within MCMC, the posterior distribution for the latent variables are proportional to $p(\boldsymbol{a}_t|\boldsymbol{a}_{t-1}, \tau^2, \sigma_0^2, \{\zeta_{\mu_j}(u_i) : j = 1, 2\}, \{\zeta_c(u_i)\}, \{\zeta_q(u_i)\})p(\{\zeta_{\mu_1}(u_i)\})$. The process will be similar for $\mu_2(s)$, $c(s)$ and $q(s)$. The prior for the latent variables $\zeta_{\mu_2}(u_i)$ are $N(0, \sigma_{\mu_2})$. The kernel convolution will be applied to the log variance, resulting in a prior of $N(0, \sigma_c)$ for $\zeta_c(u_i)$ where $\log(c(s)) = \mu_c + \sum_{i=1}^{Q} k_\zeta(u_i, s)\zeta_c(u_i)$. Using a probit transform, we can assign a Gaussian process to the

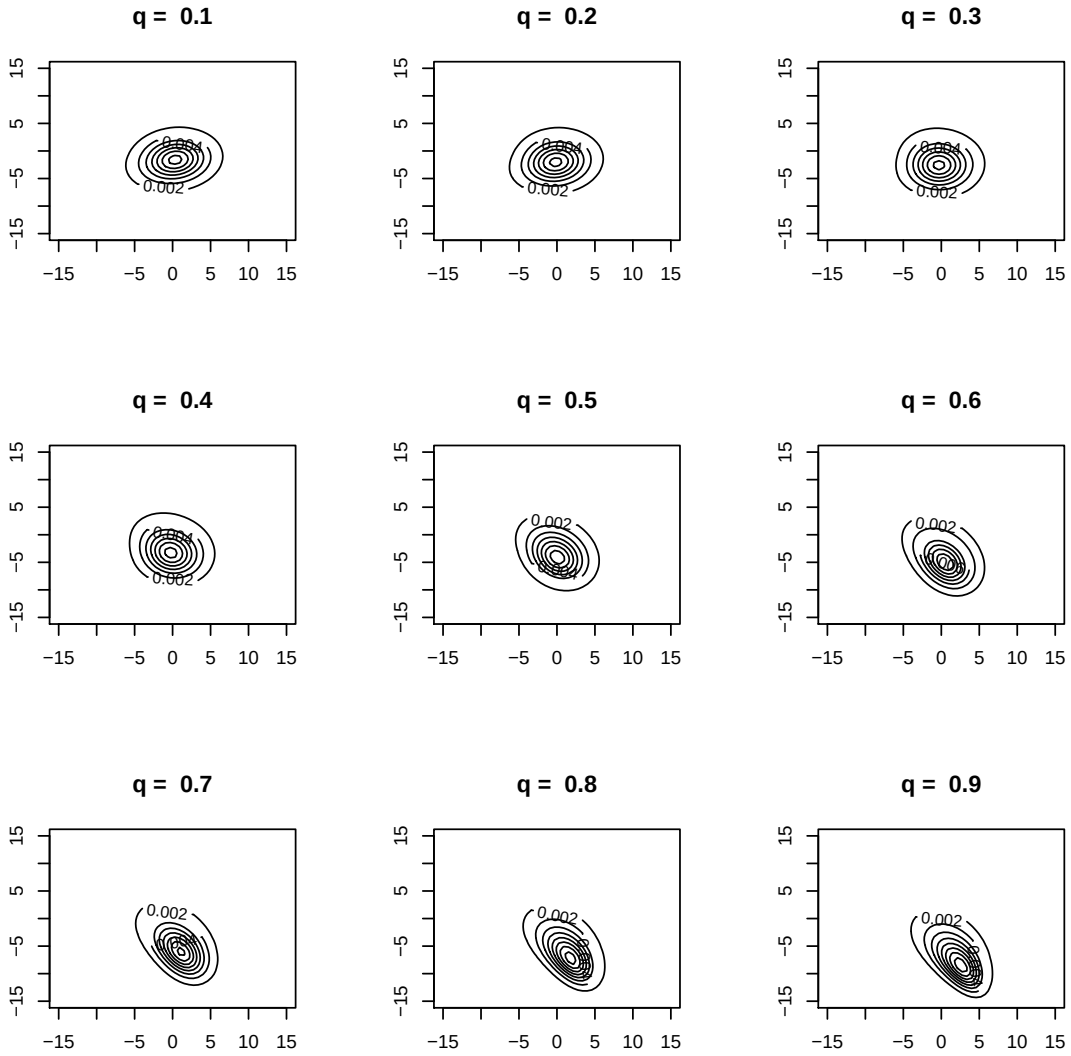


Figure 5.2: A bivariate stable density is shown with a scaled Bernstein polynomial measure and a geometric weights base distribution. $\boldsymbol{\mu} = (0, 0)'$ and $c = 2\pi$ for each plot. Only the geometric weight, q, changes. The first 10 atoms are $(2, 5.14, 3.6, .4, 3.4, .6, 3.5, .5, 3.5, .5)$. The other atoms were randomly drawn.

inverse normal CDF of the process $q(s)$. This results in $\phi^{-1}(q(s)) = \mu_q + \sum_{i=1}^{Q} k_\zeta(u_i, s)\zeta_q(u_i)$ where $\zeta_q(u_i) \sim N(0, \sigma_q)$. The parameter σ_{μ_1} is given an Inverse Gamma (IG) prior. Conditional on $\zeta_{\mu_1}(u_i)$ for $i = 1, ..., Q$, the posterior distribution for the hyperparameter will also be IG. The posterior distributions for σ_{μ_2}, σ_c, and σ_q will also be conjugate if they are paired with IG priors. With normal priors on μ_{μ_1}, μ_{μ_2}, μ_c, and μ_q, the posteriors are conjugate as well.

The parameter α for the stable distribution can be difficult to learn. To aid in estimation, we assume a discrete prior for α, between 1 and 2. With a uniform discrete prior, the posterior probability that $\alpha = a_i$ is proportional to the likelihood evaluated at $\alpha = a_i$. Another advantage of discretizing α is that the integrals in the basis coefficients from Table 5.1 can be calculated prior to the MCMC for each Bernstein polynomial and for each possible value for α, resulting in a significant speed-up. The $Z_{j,m}$ variables are also discrete. For each set $Z_{j1}, ..., Z_{jM}$, the single variable which is equal to 1 can be sampled from a discrete posterior. Allowing l_j to be an indicator variable where $l_j = m$ when Z_{jm} is 1, the probabilities are again proportional to the likelihood evaluated at $l_j = m$, which means that it can be sampled discretely with posterior probability

$$p(l_j = m | \{\boldsymbol{a}_t : t = 1, ..., T\}, \cdot) = \frac{\prod_{t=1}^{T} p(\boldsymbol{a}_t | \boldsymbol{a}_{t-1}, \cdot, l_j = m) p(l_j = m)}{\sum_{i=1}^{M} \prod_{t=1}^{T} p(\boldsymbol{a}_t | \boldsymbol{a}_{t-1}, \cdot, l_j = i) p(l_j = i)}.$$

This should be done for all J sets of latent variables. The discreteness of these variables lends itself to parallelization, as does sampling the discretized variable α. This is done by sending the calculations of the components of the discrete probabilities to different nodes and then collecting the proportional posteriors to calculate the probabilities. For a large enough data set and using M nodes, this has been seen to double the speed of the MCMC.

For this work, the parallelization was done in C++ using openMP.

5.3.3 Thresholding

The method of decomposing the process and the kernel using a basis expansion requires specification of how to truncate to a finite number of basis functions. In one dimension, the number of basis functions required for accurately representing the kernel in IDE modeling is reasonable, perhaps between 20 and 100. To achieve the same level of accuracy in 2 dimensions, hundreds of basis functions could be required. The main determining factor of the optimal number of basis functions to use is the width of the kernel compared to the range of the data. For example, when using a normal kernel, a higher variance requires fewer basis functions. This means that the optimal number of basis functions to use is not constant but changes throughout the MCMC. Recall that B_θ is a matrix where the (i, j)-th element is the j-th basis function at location s_i. Figure 5.3 shows how the width of the kernel affects the number of influential basis functions. The left plots show the kernel and the right plots show the largest basis coefficient of all the locations by basis function. Essentially, it is the column maximum of the matrix B_θ. The kernel is the stable distribution. The only difference between the kernel in the top and bottom rows is the parameter c, which doesn't affect the location or orientation, just the scale. The plots on the right show that the larger kernel requires a smaller number of basis functions. Two other observations can be made from this. The first is that a small percentage of these are significantly larger than 0. The other observations is that the size of the coefficient is not fully determined by the order of the frequencies. As basis function index increases, the frequency of the function increases. The general trend is that the coefficients get smaller,

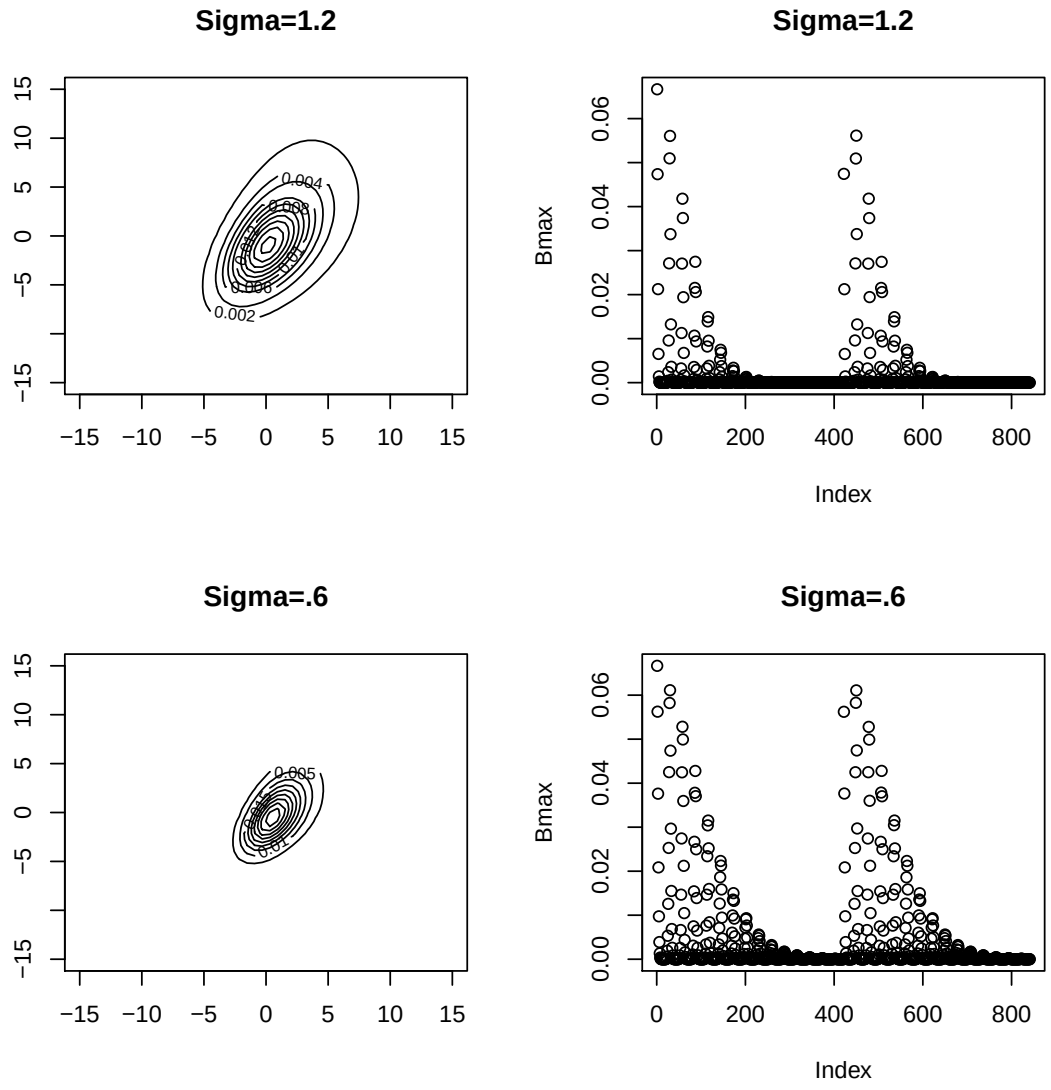


Figure 5.3: The left plots show the kernel and the right plots shows the maximum coefficient for every basis function. The first half corresponds to cosine basis coefficients and the second half corresponds to sine basis coefficients, hence the two peaks.

but several smaller frequency basis functions are not significant. By exploiting these facts the dimension of the state space can be intelligently decreased.

The MCMC can be adjusted at each iteration to include only the most important basis functions and decrease computational time of the algorithm. After calculating B_θ, the column maximum of B_θ can be used to decide the number of basis functions. There are several options for thresholding in the literature. Hard thresholding and soft thresholding are two of the most common. Hard thresholding involves setting all coefficients less than a certain value to 0, $b_j^* = b_j \times I(|b_j| > \epsilon)$. Soft thresholding additionally subtracts ϵ from values which are non-zero, $b_j^* = sign(b_j)(|b_j| - \epsilon)I(b_j > \epsilon)$. There are several arguments for and against either of these thresholding techniques. A more elegant option is the threshold based on the generalized double Pareto distribution (Armagan et al., 2013). Using this more technical approach maintains the continuity of the target function without over-shrinking. This method sets b_j to 0 when $b_j < \epsilon\sqrt{(\alpha+1)}$. When $b_j > \epsilon\sqrt{(\alpha+1)}$ then

$$b_j^* = \begin{cases} \frac{b_j - \epsilon\sqrt{(\alpha+1)} + [b_j^2 + 2b_j\epsilon\sqrt{(\alpha+1)} - 3\epsilon^2(\alpha+1)]^{1/2}}{2} & \text{if } b_j > 0 \\ \frac{b_j + \epsilon\sqrt{(\alpha+1)} - [b_j^2 + 2b_j\epsilon\sqrt{(\alpha+1)} - 3\epsilon^2(\alpha+1)]^{1/2}}{2} & \text{if } b_j < 0 \end{cases}$$

The value for ϵ defines the level of truncation and α controls the shrinkage. In order to be able to predict the length of the learning algorithm, it may help to keep the number of basis functions constant. The only way to accomplish this would be to change the thresholding level, which is ϵ in the equations above. This approach will also ensure that a reasonable number of basis functions will always be present. For this work, the generalized double Pareto thresholding was used with ϵ chosen to yield a fixed number of basis functions for each iteration of the MCMC.

5.4 SST Data Analysis

Sea surface temperature (SST) in the tropical Pacific Ocean has been useful in predicting certain phenomenon (Philander, 1985). The most prominent of these is the El Niño, which is a warming occurring between -5° and 5° Latitude and 180° and 240° E Longitude. This warming results in a shift of nutrients in the water which can affect agriculture and economy in several countries. There is a rich history of work dedicated to predicting when El Niño will occur, and its counterpart La Niña which follows. It follows a 2 to 7 year cycle and typically begins in Autumn, staying as long as a year. While many deterministic physical models have arisen to explain and predict the occurrences (Jan van Oldenborgh et al., 2005), much success has come from simply using Sea Surface temperature data over a large region in the Pacific Ocean in a stochastic model. These methods include linear systems (Penland and Magorian, 1993), but nonlinear methods have proven more successful (Wikle and Hooten, 2010; Cressie and Wikle, 2011). Non-linear methods have been applied to this exact SST data in the context of IDE models (Wikle and Holan, 2011), although the specific nature of the kernel distribution was not the focus of that application.

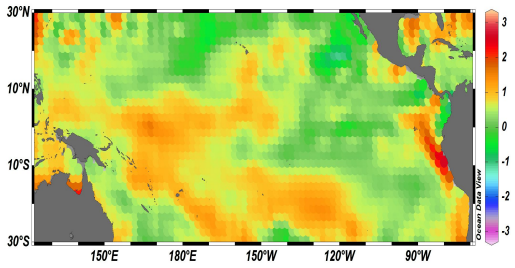
We will illustrate the bivariate stable kernel for IDE modeling using SST data. The data includes 2261 locations on a 2° by 2° resolution grid. It is collected monthly from January 1970 to March 2003. The anomalies of 9 months of this data is shown in Figure 5.4. It is known that a mild El Niño began in late summer of 2002. It can be seen in the figure how the warm temperatures gather near the equator East of the Date Line, indicating El Niño. We will compare the spatially varying stable kernel with the spatially varying Gaussian kernel to see how effective it is in prediction, and specifically to see how

the models perform when predicting the 2002 El Niño phenomenon months in advance.

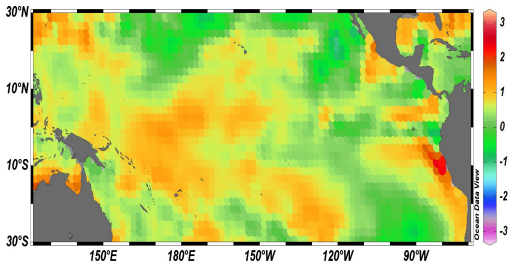
We apply the model in equations (5.2) - (5.6) to the SST monthly anomalies, substituting in equations (5.7) and (5.8) when appropriate. The parameter space is quite large, including the observational variance σ^2, the process variance τ^2 and the kernel parameters, which include the latent variables involved in the kernel convolution for the processes $\boldsymbol{\mu}(\boldsymbol{s})$, $q(\boldsymbol{s})$, and $c(\boldsymbol{s})$, and the hyperparameters for the latent ζ parameters for each process. The locations of the data are given in latitude and longitude, but for the analysis they are scaled to between -10 and 10 in both directions and then scaled back to the original locations for inference. The full model is:

$$
\begin{aligned}
\boldsymbol{Y}_t|\boldsymbol{a}_t,\sigma^2 &\sim N(\boldsymbol{\Psi}\boldsymbol{a}_t,\sigma^2\boldsymbol{I}_t), \quad t=1,...,T, \quad \sigma^2\sim \text{IG}(\alpha_\sigma,\beta_\sigma)\\
\boldsymbol{a}_t|\boldsymbol{a}_{t-1},\tau^2,\boldsymbol{\theta} &\sim N(\boldsymbol{G}\boldsymbol{B}_\theta\boldsymbol{a}_{t-1},\tau^2\boldsymbol{G}\boldsymbol{V}\boldsymbol{G}'), \quad \tau^2\sim \text{IG}(\alpha_\tau,\beta_\tau)\\
\gamma(z) &= c(\boldsymbol{s})\sum_{k=1}^{M} w_{M,m}\text{Beta}(z/2\pi|m,M-m+1)/2\pi, \quad \alpha\sim p(\alpha)\\
w_{M,m} &= \sum_{j=1}^{J} q(\boldsymbol{s})(1-q(\boldsymbol{s}))^{j-1}Z_{jm}, \quad (Z_{j1},...,Z_{jM})\sim \text{MN}(1,(1/M,...,1/M))\\
\phi^{-1}(q(\boldsymbol{s})) &= \mu_q\mathbf{1}+\boldsymbol{K}_\zeta\boldsymbol{\zeta_q}, \quad \zeta_q(u_i)\sim N(0,\sigma_q), \quad \mu_q,\sigma_q\sim p(\mu_q)p(\sigma_q)\\
\log c(\boldsymbol{s}) &= \mu_c\mathbf{1}+\boldsymbol{K}_\zeta\boldsymbol{\zeta_c}, \quad \zeta_c(u_i)\sim N(0,\sigma_c), \quad \mu_c,\sigma_c\sim p(\mu_c)p(\sigma_c)\\
\mu_i(\boldsymbol{s}) &= \mu_{\mu_i}\mathbf{1}+\boldsymbol{K}_\zeta\boldsymbol{\zeta_{\mu_i}}, \quad \zeta_{\mu_i}(u_i)\sim N(0,\sigma_{\mu_i}), \quad \mu_{\mu_i},\sigma_{\mu_i}\sim p(\mu_{\mu_i})p(\sigma_{\mu_i})
\end{aligned}
$$

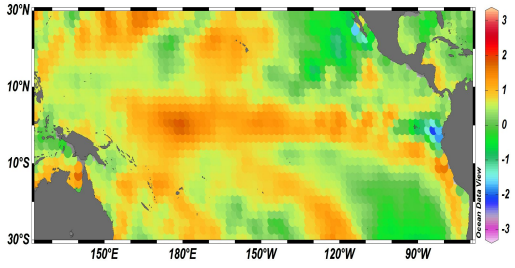
The construction of $\boldsymbol{B}_\theta$, $\boldsymbol{\Psi}$, and G are detailed in section 2.3. The matrix $\boldsymbol{V}$ is the unscaled spatial covariance matrix based on a Matern covariance function with $\kappa = 1.5$ and an effective range of 2. The number of Bernstein polynomials used is $M = 40$. The priors for σ^2 and τ^2 are IG(5, 3). The matrix $\boldsymbol{K}_\zeta$ maps the latent $\boldsymbol{\zeta}_\theta = (\zeta_\theta(u_1),...,\zeta_\theta(u_Q))'$ vectors for $\theta \in \{q,\sigma,\mu_1,\mu_2\}$ to the processes governing the IDE kernel parameters. The values of $\boldsymbol{K}_\zeta$



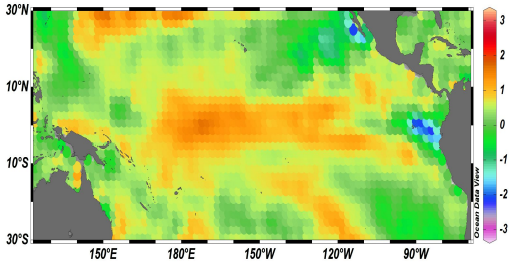
(a) April 2002



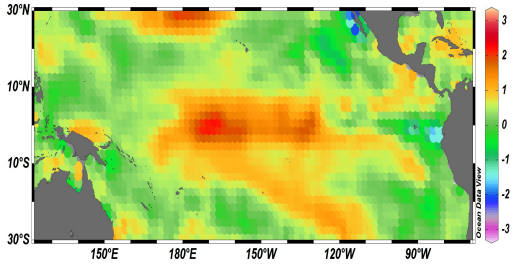
(b) May 2002



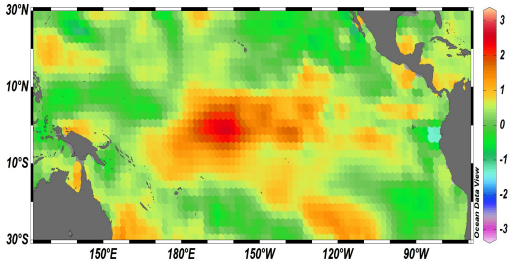
(c) June 2002



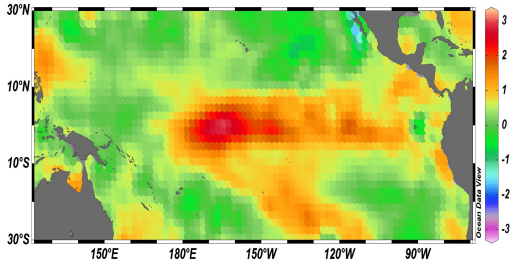
(d) July 2002



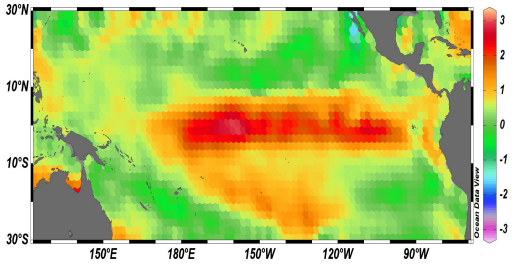
(e) August 2002



(f) September 2002



(g) October 2002



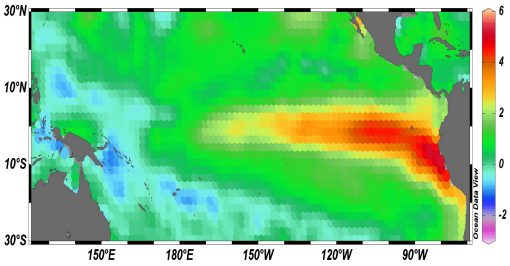
(h) November 2002

Figure 5.4: Data is shown for sea surface temperature anomalies from April to December 2002.

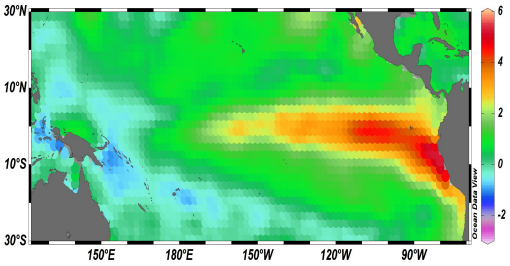
correspond to convolution kernel where the (i, j)-th element is $k_\zeta(s_i - u_j)$. The convolution kernel is a Matern function with $\kappa = 2.5$ and an effective range of 4 for all the parameters, forcing a smooth evolution of the processes across the domain. The knots are chosen on a 20 by 20 grid from -11 to 11 in both directions, resulting in a dimension reduction of the spatially varying parameter sets from 2261 to 400. The priors for the means of these processes are $\mu_q \sim N(-1, .5)$, $\mu_c \sim N(0, 4)$, and μ_{μ_1}, $\mu_{\mu_2} \sim N(0, 1)$. The scale terms for the process covariances are given priors of IG$(4, 3)$ for σ_c and σ_{μ_i} and IG$(10, 6)$ for σ_q. These priors are based on the scale of the data and reasonable shapes of kernels, but aren't too restrictive. The process $q(\boldsymbol{s})$ is very sensitive to these priors, as the posterior is not necessarily identifiable, especially for very diffuse priors. There may be several combinations of $q(\boldsymbol{s})$ and the latent Z_{jk} variables which results in the same values for $w_{M,m}$, which should be identifiable. The other parameters are not overly sensitive to the prior.

The model is fit using MCMC methods described in 2.4. 75,000 samples were taken of the posterior distributions with the first 60,000 as burn-in, leaving 15,000 samples. The convergence was checked using trace plots of the values of the kernel densities at each location. That is to say that the trace plots suggested convergence by 60,000 iterations of the values for $k(\boldsymbol{s}|\boldsymbol{\mu}(\boldsymbol{s}), \sigma(\boldsymbol{s}), q(\boldsymbol{s}))$ for various locations throughout the domain of the data. The means of the in-sample posterior predictive distributions of the data for the months leading up to the 1997 El Niño are shown in Figure 5.5. Similar accuracy can be seen for all time points.

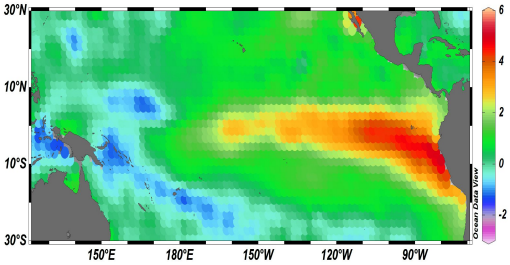
The posterior kernel and associated measure may reveal information about the nature of dependence between locations. In section 5.2.2, the property of elliptical symmetry



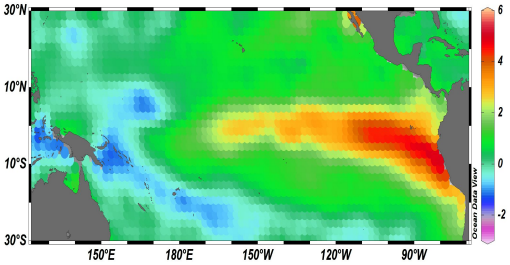
(a) Data June 1997



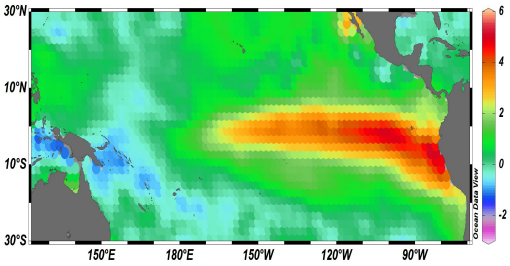
(b) Fitted June 1997



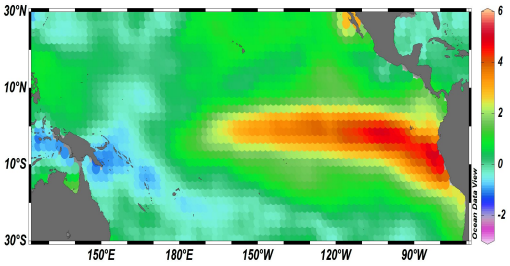
(c) Data July 1997



(d) Fitted July 1997



(e) Data August 1997



(f) Fitted August 1997

Figure 5.5: Data and fitted values are compared for the months leading up to El Niño in 1997. The fitted values are the means of the in-sample posterior predictive distributions.

is connected to the values of γ. The posterior kernel for 3 locations is shown in Figure 5.6 with the associated densities scaled by 2π. It is hard to detect differences by eye between the kernels at the different locations. The differences in the measure are much more easily detected and are clearly different for these three locations.

To detect trends, we devise a method of measuring symmetry across the spatial field. Lemma 6 states that elliptical symmetry is equivalent to the property that $\gamma(z) = \gamma(z + \pi)$ for all z. We can create a metric of elliptical using symmetry

$$sym_{ell}(\boldsymbol{s}) = \int_0^{\pi} (\gamma_{\boldsymbol{s}}(z) - \gamma_{\boldsymbol{s}}(z + \pi))^2 dz.$$

Similarly, a metric of spherical symmetry can be defined as

$$sym_{sph}(\boldsymbol{s}) = \int_0^{2\pi} (\gamma_{\boldsymbol{s}}(z)/c(\boldsymbol{s}) - (2\pi)^{-1})^2 dz.$$

These symmetry metrics are plotted by the spatial location in Figure 5.7. The kernels at locations north of the equator are symmetric whereas the kernels for locations south of the equator are non-symmetric. Additionally, the spherical symmetry map is very similar to the elliptical symmetry map.

We will assess model performance through its predictive power. An additional model fit was conducted based on a spatially varying Gaussian kernel, following Wikle (2002) and Xu et al. (2005). We score the one-step ahead predictions for 399 time points based on energy scores from equation (3.2). We find that the stable kernel IDE model scores better in 229 time points, which is 57.4% of all time points. Based on these alone, the advantage seems minimal. However slight the departures from elliptical symmetry were, they do exist, but this does not seem to affect the scoring results. We also compare the models using

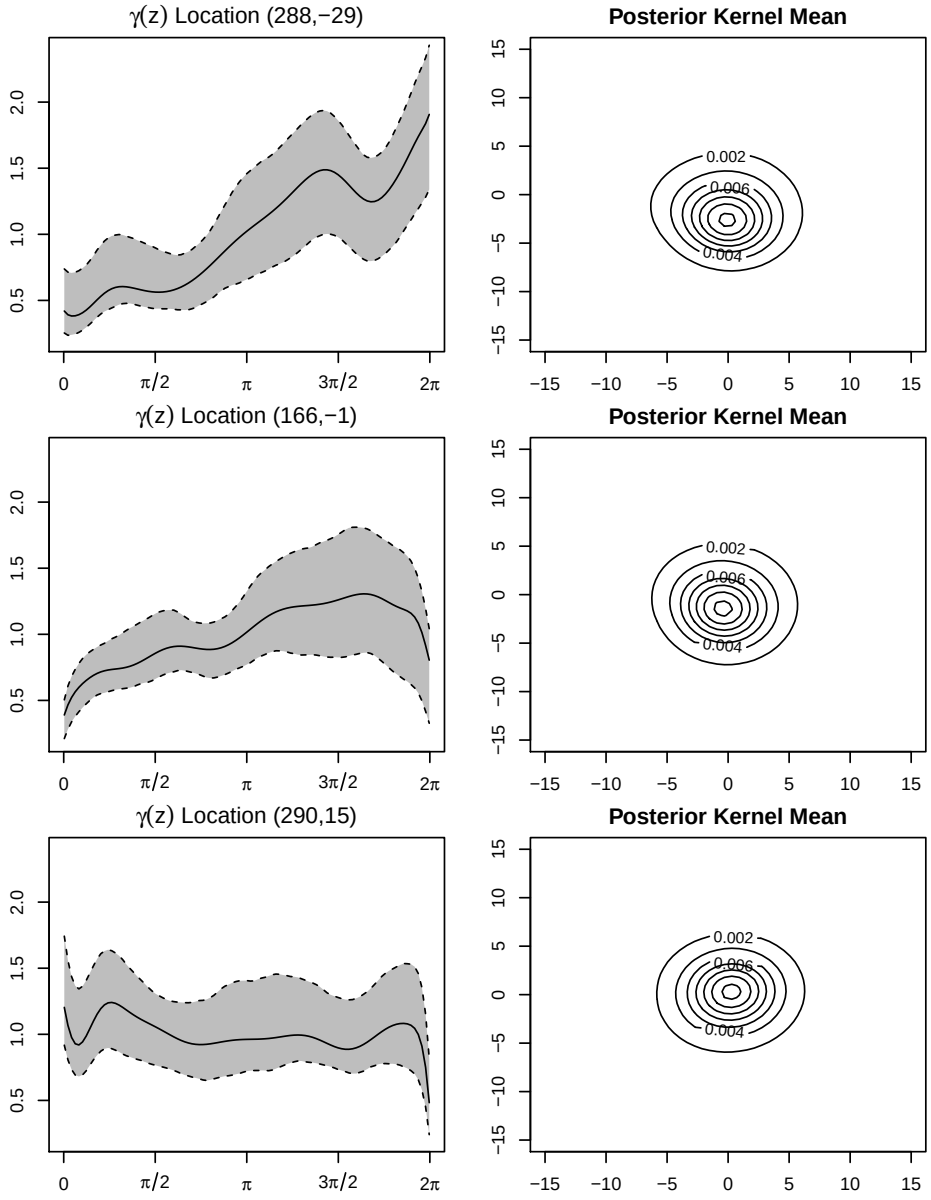
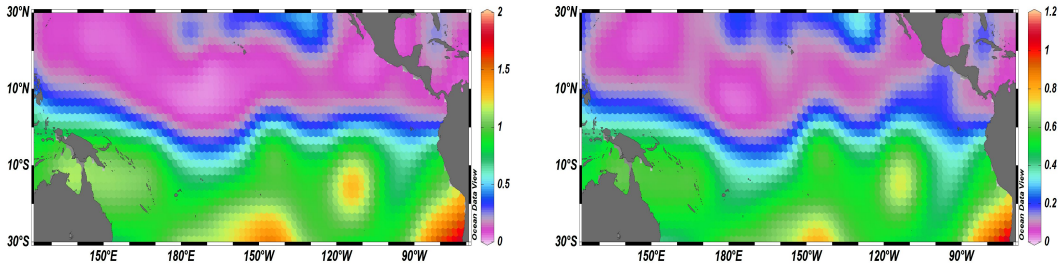


Figure 5.6: The scaled density corresponding to the measure Γ and associated 95% credible bands are shown for three locations on the left with the associated posterior mean kernel on the right.



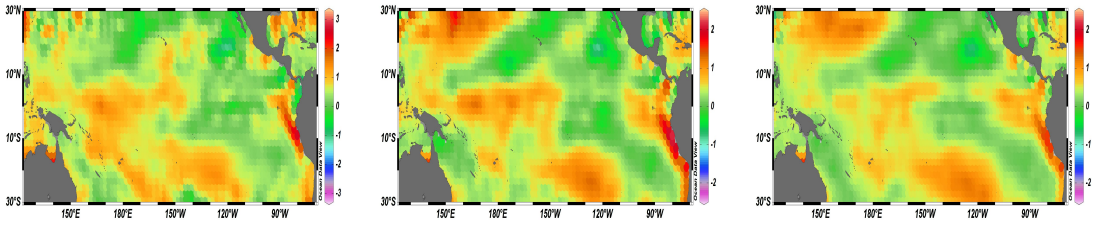
(a) Elliptical Symmetry

(b) Spherical Symmetry

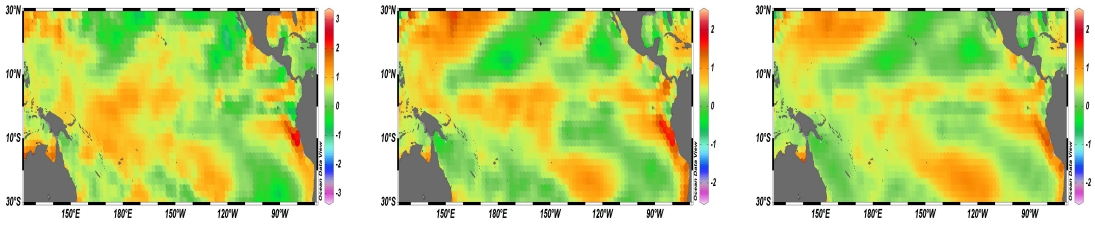
Figure 5.7: The symmetry metrics are shown across the spatial field for both elliptical and spherical symmetry. For both metrics, smaller values are associated with more symmetric kernels.

K-step ahead forecasts. This is done by propagating the state variables through the process level of the model, $\boldsymbol{a}_t^* \sim N(\boldsymbol{G}\boldsymbol{B}_\theta \boldsymbol{a}_{t-1}, \tau^2 \boldsymbol{G}\boldsymbol{V}\boldsymbol{G}')$. This can be propagated several steps ahead followed by drawing the prediction $Y_t^* \sim N(\boldsymbol{\Psi}\boldsymbol{a}_t^*, \sigma^2 \boldsymbol{I})$. To draw from the posterior K-step ahead predictive distribution, the last year of data was left out of the analysis, resulting in 387 time points ending in March 2002 being included in the model fit. The posterior distributions for $Y_{388}^*, ..., Y_{399}^*$ will be drawn as part of the MCMC. The means of these posterior distributions are shown in Figure 5.8 for the stable and normal kernel IDE models compared against the truth for April through July 2002. Figure 5.9 shows the same for the months August through November 2002. The real data shows the El Niño which is known to have occurred in 2002. Both the predictions include the El Niño warming to some degree, but it is clear that the intensity of the predicted warming using the stable kernel IDE model is much closer to the truth than the normal kernel IDE model.

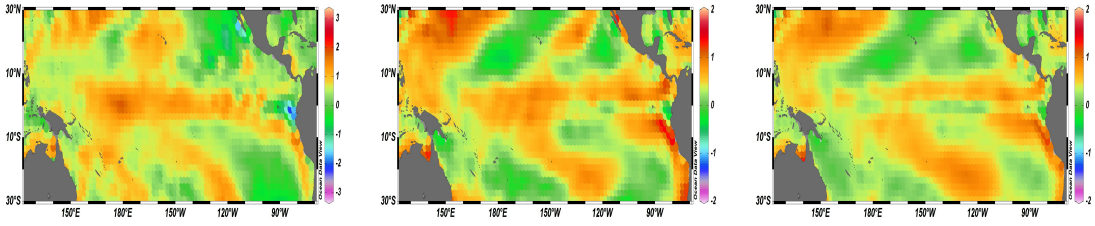
To assign specific values to these predictions, we again use energy scores. The



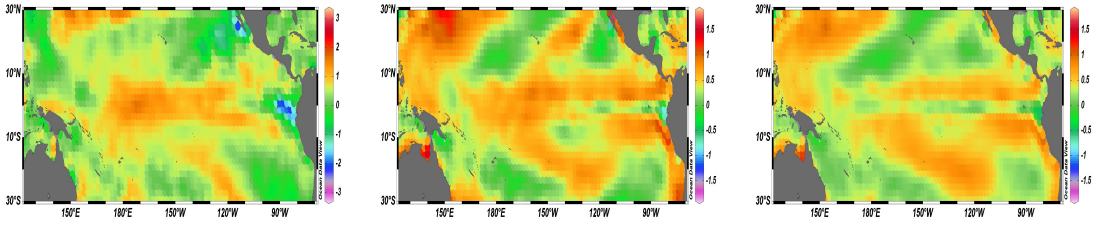
(a) April 2002 Truth (b) April 2002 Stable Predictions (c) April 2002 Normal Predictions



(d) May 2002 Truth (e) May 2002 Stable Predictions (f) May 2002 Normal Predictions

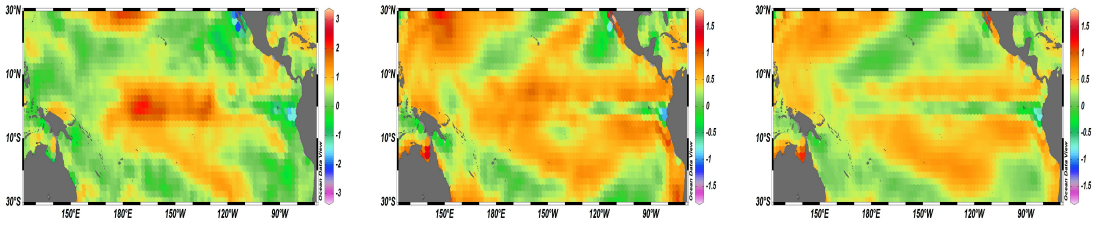


(g) June 2002 Truth (h) June 2002 Stable Predictions (i) June 2002 Normal Predictions

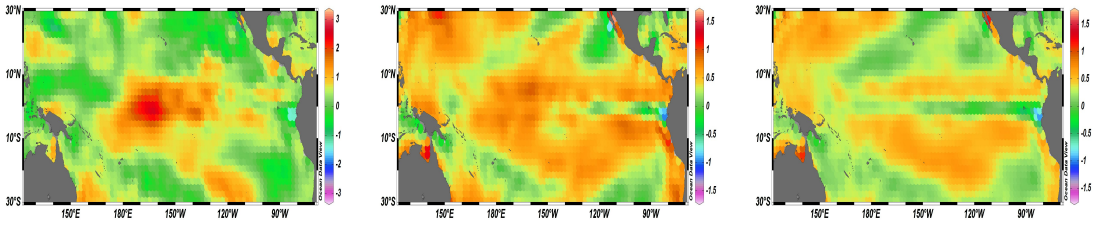


(j) July 2002 Truth (k) July 2002 Stable Predictions (l) July 2002 Normal Predictions

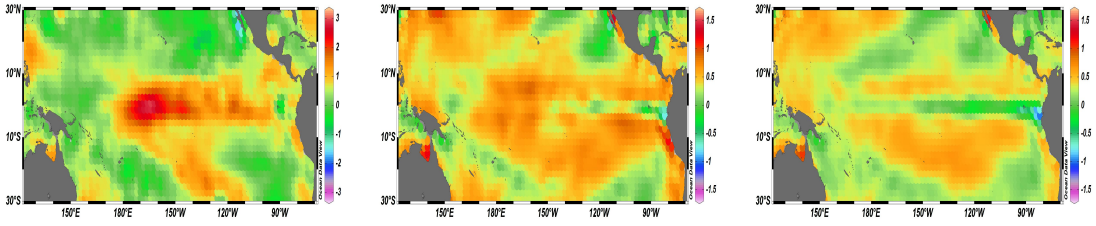
Figure 5.8: The data and posterior K-step ahead predictions using the stable and normal kernels are shown given information through March 2002. The months April 2002 through July 2002 are shown.



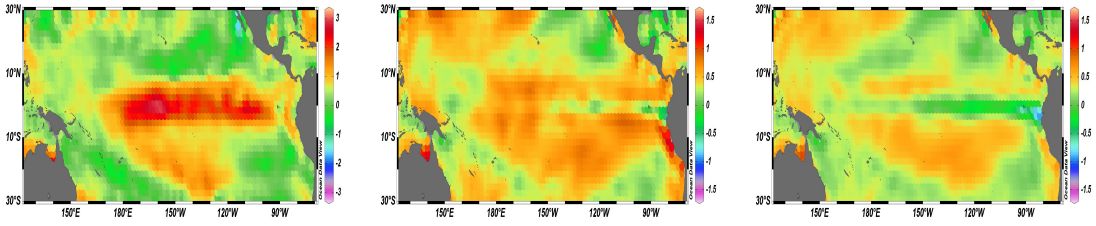
(a) Aug 2002 Truth (b) Aug 2002 Stable Predictions (c) Aug 2002 Normal Predictions



(d) Sept 2002 Truth (e) Sept 2002 Stable Predictions (f) Sept 2002 Normal Predictions



(g) Oct 2002 Truth (h) Oct 2002 Stable Predictions (i) Oct 2002 Normal Predictions



(j) Nov 2002 Truth (k) Nov 2002 Stable Predictions (l) Nov 2002 Normal Predictions

Figure 5.9: The data and posterior K-step ahead predictions using the stable and normal kernels are shown given information through March 2002. The months August 2002 through November 2002 are shown.

energy scores for each prediction is shown in Figure 5.10. The very first step is the only step where the Gaussian is lower. The stable kernel IDE model scores better for all steps after 1. The effect seems to be diminishing after 12 steps, which should be expected. This strongly supports the stable kernel IDE model to fit this data over the Gaussian kernel.

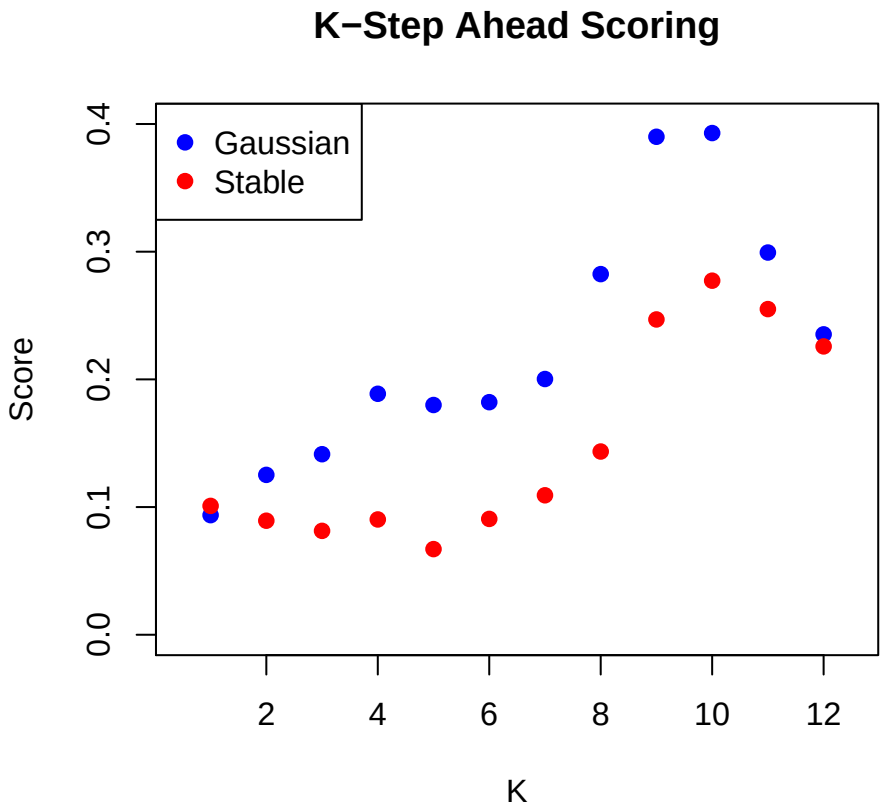


Figure 5.10: Scoring for the K-step ahead predictions. The first 9 of these are the scores for the panels shown in Figures 5.8 and Figure 5.9 for the stable and Gaussian respectively.

There are several ways to numerically declare an El Niño event. Most of these involve high SST anomalies in certain regions in the Pacific. For example, the official National Oceanic and Atmospheric Administration (NOAA) criterion involves the block

average SST anomalies in El Niño region 3.4 to be above .5° C for 3 consecutive months, although this is modified to 5 consecutive months for the NOAA's Climate Prediction Center (Larkin and Harrison, 2005). The El Niño region 3.4 includes 120° to 170° W Longitude and -5° to 5° Latitude. Figure 5.11 shows the block averages for the data and fitted models. These estimates are sampled from the posterior predictive distributions for K-steps ahead given all information through March 2002. While the point estimates severely undershoot the data for both models, the stable kernel IDE model contains the truth in every credible band whereas the normal kernel IDE model misses in several months. Also, by the NOAA definition of an El Niño occurrence, the point estimate for the stable kernel IDE model would have predicted an El Niño whereas the normal kernel IDE model would not have. In fact, out of 15,000 samples, the normal kernel IDE model predicted a 55% chance of an El Niño event for a 3-month criterion and a 20% chance for the 5 month criterion. The stable kernel IDE model predicted a 76% chance of an El Niño event for the 3-month criterion and 37% chance for th 5 month criterion.

5.5 Summary

The bivariate stable kernel has been proposed as an alternative to the normal kernel in spatio-temporal IDE modeling. Using the normal kernel as proposed in Xu et al. (2005) results in 5 spatial processes for the parameters. One of the advantages of the bivariate stable kernel as proposed in this chapter is that it provides more flexibility than the Gaussian kernel with only 4 spatial processes on the parameters. Using a geometric weights prior for the Bernstein polynomials allows a great deal of flexibility, but could be

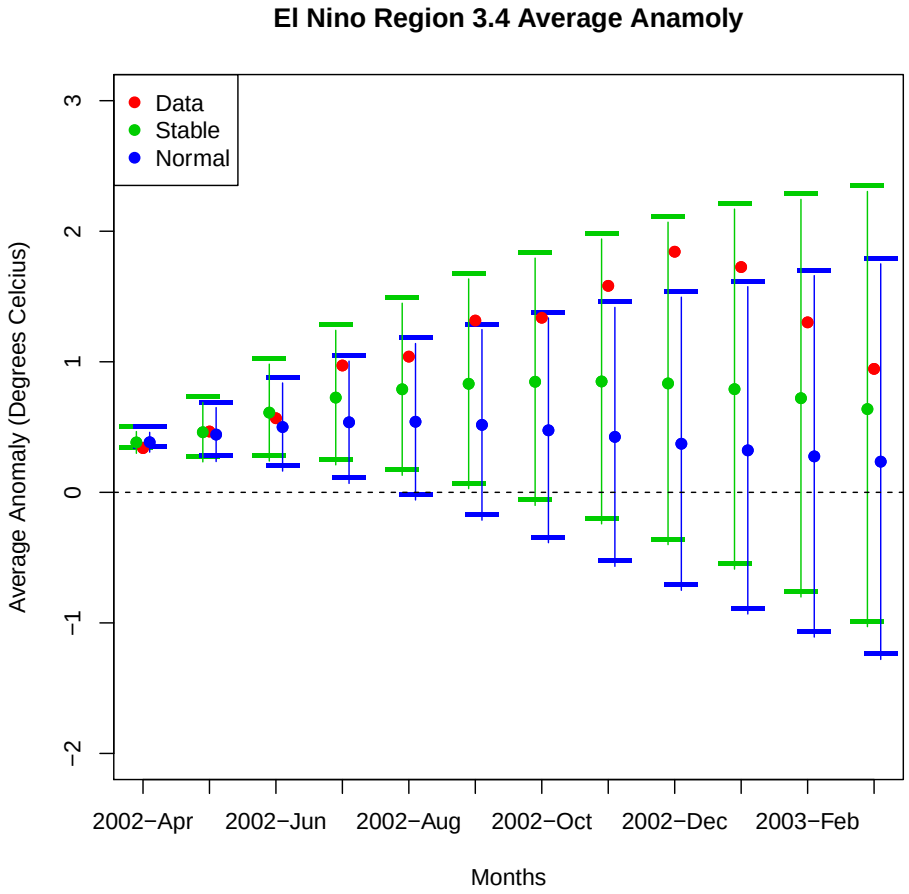


Figure 5.11: Block average SST anomalies for the El Niño regions 3.4 for the data are shown, as well as the estimated posterior predictive means for the IDE model with the stable and Gaussian kernel. 95% credible bands are shown for the model fits as well.

extended even more. For example, Petrone (1999) uses the Dirichlet process for the base distribution controlling the weights. The flexibility comes with the cost of ease in estimation. Using the SST data, we have shown the value of using these models. Prediction improves, especially for more than one step ahead.

Chapter 6

Conclusion

There are several ways to extend IDE modeling which were not discussed in this dissertation. Polynomial IDE modeling is discussed in Wikle and Holan (2011). Interactions between different locations of the process at previous time points in the model contribute to the process at the current time point. While it is possible to assign a physical interpretation of the process to characteristics of the kernel in linear IDE modeling, the interpretations of the kernel contribution in the polynomial IDE case is not clear. Despite the lack of interpretability, the result is a non-linear model, which may be more appropriate than linear modeling in several cases. In fact, Wikle and Holan (2011) compares polynomial IDE modeling with linear modeling using the SST data from Chapter 5 and concludes that the model improves in several measurable ways when using non-linear models. While using a non-parametric kernel in an IDE model is very flexible, the relationship between $\boldsymbol{Y}_t$ and $\boldsymbol{Y}_{t-1}$ is still linear. It may be possible to define the kernel in some way to induce a non-linear model without the interactions. It may also be possible to apply more flexible kernels to

polynomial IDE modeling. Kernel estimation in linear IDE modeling is difficult as it is. Trying to learn flexible kernels defining the interaction may be impossible, but also may be an area of expansion for this topic.

Non-Gaussian kernels have been the focus of this dissertation, but all the error functions are assumed to be Gaussian. Another extension is to allow the error function to be more flexible. The IDE model becomes a state space model, and methods to fit non-Gaussian state-space models have been studied. For example, Kitagawa (1996) details filtering and smoothing for high-dimensional non-Gaussian and non-linear state space models. For many data sets, the assumption of Gaussianity may be restrictive for the variances. Estimation for complicated kernels has been achieved in this dissertation and estimating non-Gaussian state space models is in the literature, but computational issues may arise from combining the methods.

Even without these extensions, we have provided powerful motivation for extending the Gaussian kernel IDE models to non-Gaussian kernels. We have detailed practical ways to fit such models and demonstrated how results can be produced. We have provided an array of options, including models which are slightly more complicated to models which have full kernel flexibility. Due to the complicated way the kernels are embedded in the IDE model, much of this dissertation has been dedicated to computational methods, such as Hermite Polynomials, Hamiltonian Monte Carlo, and thresholding the basis coefficients. By using these proposed methods, flexible IDE modeling may be possible for general spatio-temporal modeling.

Bibliography

Abdul-Hamid, H. and Nolan, J. P. (1998), "Multivariate stable densities as functions of one dimensional projections," *Journal of multivariate analysis*, 67, 80–89.

Antoniak, C. E. et al. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The annals of statistics*, 2, 1152–1174.

Armagan, A., Dunson, D. B., and Lee, J. (2013), "Generalized double Pareto shrinkage," *Statistica Sinica*, 23, 119.

Brown, P. E., Roberts, G. O., Kåresen, K. F., and Tonellato, S. (2000), "Blur-generated non-separable space–time models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 847–860.

Byczkowski, T., Nolan, J. P., and Rajput, B. (1993), "Approximation of multidimensional stable densities," *Journal of Multivariate Analysis*, 46, 13–31.

Courant, R. and Hilbert, D. (1966), *Methods of mathematical physics*, vol. 1, CUP Archive.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: John Wiley & Sons.

Cressie, N. and Huang, H.-C. (1999), "Classes of nonseparable, spatio-temporal stationary covariance functions," *Journal of the American Statistical Association*, 94, 1330–1339.

Cressie, N. and Wikle, C. K. (2011), *Statistics for spatio-temporal data*, New York: John Wiley & Sons.

Dewar, M., Scerri, K., and Kadirkamanathan, V. (2009), "Data-driven spatio-temporal modeling using the integro-difference equation," *Signal Processing, IEEE Transactions on*, 57, 83–91.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007), "Generalized spatial Dirichlet process models," *Biometrika*, 94, 809–825.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The annals of statistics*, 209–230.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Gneiting, T. (2002), "Nonseparable, stationary covariance functions for space–time data," *Journal of the American Statistical Association*, 97, 590–600.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008), "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds," *Test*, 17, 211–235.

Hamilton, J. D. (1994), *Time series analysis*, vol. 2, Cambridge University Press.

Heck, W. W., Cure, W. W., Rawlings, J. O., Zaragoza, L. J., Heagle, A. S., Heggestad, H. E., Kohut, R. J., Kress, L. W., and Temple, P. J. (1984), "Assessing impacts of ozone on agricultural crops: II. Crop yield functions and alternative exposure statistics," *Journal of the Air Pollution Control Association*, 34, 810–817.

Heine, V. (1955), "Models for two-dimensional stationary stochastic processes," *Biometrika*, 42, 170–178.

Higdon, D. (1998), "A process-convolution approach to modelling temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190.

Hooten, M. B. and Wikle, C. K. (2008), "A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove," *Environmental and Ecological Statistics*, 15, 59–70.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96.

Jan van Oldenborgh, G., Balmaseda, M. A., Ferranti, L., Stockdale, T. N., and Anderson, D. L. (2005), "Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years?" *Journal of climate*, 18, 3240–3249.

Jones, R. H. and Zhang, Y. (1997), "Models for continuous stationary space-time processes," in *Modelling longitudinal and spatially correlated data*, eds. Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek-Cohen, E., Warren, W. G., and Wolfinger, R. D., Springer, pp. 289–298.

Kitagawa, G. (1996), "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *Journal of computational and graphical statistics*, 5, 1–25.

Kot, M., Lewis, M. A., and van den Driessche, P. (1996), "Dispersal data and the spread of invading organisms," *Ecology*, 77, 2027–2042.

Kottas, A., Duan, J. A., and Gelfand, A. E. (2008), "Modeling disease incidence data with spatial and spatio temporal Dirichlet process mixtures," *Biometrical Journal*, 50, 29–42.

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001), *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Boston: Birkhäuser.

Larkin, N. K. and Harrison, D. (2005), "On the definition of El Niño and associated seasonal average US weather anomalies," *Geophysical Research Letters*, 32.

Ma, C. (2003), "Nonstationary covariance functions that model space–time interactions," *Statistics & probability letters*, 61, 411–419.

MacEachern, S. N. (2000), "Dependent dirichlet processes," *Unpublished manuscript, Department of Statistics, The Ohio State University*.

Matheron, G. (1963), "Principles of geostatistics," *Economic geology*, 58, 1246–1266.

Matsui, M. and Takemura, A. (2009), "Integral representations of one-dimensional projections for multivariate stable densities," *Journal of Multivariate Analysis*, 100, 334–344.

Mena, R. H., Ruggiero, M., and Walker, S. G. (2011), "Geometric stick-breaking processes

for continuous-time Bayesian nonparametric modeling," *Journal of Statistical Planning and Inference*, 141, 3217–3230.

Neal, R. M. (2011), "MCMC using Hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, 2.

Neubert, M. G., Kot, M., and Lewis, M. A. (1995), "Dispersal and pattern formation in a discrete-time predator-prey model," *Theoretical Population Biology*, 48, 7–43.

Nolan, J. (2003), *Stable distributions: models for heavy-tailed data*, New York: Birkhauser.

Nolan, J. P. (2013), "Multivariate elliptically contoured stable distributions: theory and estimation," *Computational Statistics*, 28, 2067–2089.

— (2014), "Financial modeling with heavy-tailed stable distributions," *Wiley Interdisciplinary Reviews: Computational Statistics*, 6, 45–55.

Nolan, J. P., Gonzalez, J. G., and Nunez, R. C. (2010), "Stable filters: A robust signal processing framework for heavy-tailed noise," in *Radar Conference, 2010 IEEE*, IEEE, pp. 470–473.

Nolan, J. P., Panorska, A. K., and McCulloch, J. H. (2001), "Estimation of stable spectral measures," *Mathematical and Computer Modelling*, 34, 1113–1122.

Ogata, H. (2013), "Estimation for multivariate stable distributions with generalized empirical likelihood," *Journal of Econometrics*, 172, 248–254.

Olver, F. W. (2010), *NIST handbook of mathematical functions*, Cambridge University Press.

Panorska, A. K. (1996), "Generalized stable models for financial asset returns," *Journal of computational and applied mathematics*, 70, 111–114.

Penland, C. and Magorian, T. (1993), "Prediction of Nino 3 sea surface temperatures using linear inverse modeling," *Journal of Climate*, 6, 1067–1076.

Petrone, S. (1999), "Bayesian density estimation using Bernstein polynomials," *Canadian Journal of Statistics*, 27, 105–126.

Philander, S. (1985), "El Niño and La Niña," *Journal of the Atmospheric Sciences*, 42, 2652–2662.

Pivato, M. and Seco, L. (2003), "Estimating the spectral measure of a multivariate stable distribution via spherical harmonic analysis," *Journal of Multivariate Analysis*, 87, 219–240.

Prado, R. and West, M. (2010), *Time series: modeling, computation, and inference*, Florida: CRC Press.

Richardson, R., Kottas, A., and Sansó, B. (2015), "Flexible Integro-Difference Equations for Spatio-Temporal Modeling," *Journal of Agricultural, Biological, and Environmental Sciences*.

Robeson, S. and Steyn, D. (1990), "Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations," *Atmospheric Environment. Part B. Urban Atmosphere*, 24, 303–312.

Salas-Gonzalez, D., Kuruoglu, E. E., and Ruiz, D. P. (2009), "Modelling and assessing dif-

ferential gene expression using the alpha stable distribution," *The International Journal of Biostatistics*, 5.

— (2010), "Modelling with mixture of symmetric stable distributions using Gibbs sampling," *Signal Processing*, 90, 774–783.

Samorodnitsky, G. and Taqqu, M. S. (1997), "Stable Non-Gaussian Random Processes," *Econometric Theory*, 13, 133–142.

Scerri, K., Dewar, M., and Kadirkamanathan, V. (2009), "Estimation and model selection for an IDE-based spatio-temporal model," *Signal Processing, IEEE Transactions on*, 57, 482–492.

Schmidt, A. M. and O'Hagan, A. (2003), "Bayesian inference for non-stationary spatial covariance structure via spatial deformations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 743–758.

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

Shumway, R. H. and Stoffer, D. S. (2011), *Time series analysis and its applications: with R examples*, New York: Springer.

Sigrist, F., Künsch, H. R., and Stahel, W. A. (2012), "A Dynamic Nonstationary Spatio-temporal Model for Short Term Prediction of Precipitation," *The Annals of Applied Statistics*, 6, 1452–1477.

Smith, B. J. (2007), "boa: an R package for MCMC output convergence assessment and posterior inference," *Journal of Statistical Software*, 21, 1–37.

Stein, M. L. (2005a), "Nonstationary spatial covariance functions," *Unpublished technical report. Available.*

— (2005b), "Space–time covariance functions," *Journal of the American Statistical Association*, 100, 310–321.

Steutel, F. W. and Harn, K. V. (2003), *Infinite divisibility of probability distributions on the real line*, Florida: CRC Press.

Storvik, G., Frigessi, A., and Hirst, D. (2002), "Stationary space-time Gaussian fields and their time autoregressive representation," *Statistical Modelling*, 2, 139–161.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, New York: Springer Verlag, 2nd ed.

Whittle, P. (1986), *Systems in stochastic equilibrium*, New York: John Wiley & Sons, Inc.

Wikle, C. K. (2002), "A kernel-based spectral model for non-Gaussian spatio-temporal processes," *Statistical Modelling*, 2, 299–314.

Wikle, C. K. and Cressie, N. (1999), "A dimension-reduced approach to space-time Kalman filtering," *Biometrika*, 86, 815–829.

Wikle, C. K. and Holan, S. H. (2011), "Polynomial nonlinear spatio-temporal integro-difference equation models," *Journal of Time Series Analysis*, 32, 339–350.

Wikle, C. K. and Hooten, M. B. (2010), "A general science-based framework for dynamical spatio-temporal models," *Test*, 19, 417–451.

Xu, K., Wikle, C. K., and Fox, N. I. (2005), "A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities," *Journal of the American Statistical Association*, 100, 1133–1144.