CARNEGIE MELLON UNIVERSITY

HIGH-DIMENSIONAL ADAPTIVE BASIS DENSITY ESTIMATION

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY
In
STATISTICS

by

Susan Buchman

Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

May, 2011

Abstract

In the realm of high-dimensional statistics, regression and classification have received much attention, while density estimation has lagged behind. Yet there are compelling scientific questions which can only be addressed via density estimation using high-dimensional data, such as the paths of North Atlantic tropical cyclones. If we cast each track as a single high-dimensional data point, density estimation allows us to answer such questions via integration or Monte Carlo methods. In this dissertation, I present three new methods for estimating densities and intensities for high-dimensional data, all of which rely on a technique called diffusion maps. This technique constructs a mapping for high-dimensional, complex data into a low-dimensional space, providing a new basis that can be used in conjunction with traditional density estimation methods. Furthermore, I propose a re-ordering of importance sampling in the high-dimensional setting. Traditional importance sampling estimates high-dimensional integrals with the aid of an instrumental distribution chosen specifically to minimize the variance of the estimator. In many applications, the integral of interest is with respect to an estimated density. I argue that in the high-dimensional realm, performance can be improved by reversing the procedure: instead of estimating a density and then selecting an appropriate instrumental distribution, begin

with the instrumental distribution and estimate the density with respect to it directly. The variance reduction follows from the improved density estimate. Lastly, I present some initial results in using climatic predictors such as sea surface temperature as spatial covariates in point process estimation.

Acknowledgements

This thesis is entirely due to the guidance and insight of my advisors, Chad Schafer and Ann Lee. Not only have they been excellent statistical and professional role models, but I would wager that in the history of higher education, no advisors have ever responded with more kindness, understanding, and patience to the news that a student was expecting twins. Thanks to Peter Freeman for his deep involvement with this work and our many hours spent in meetings helping to improve it. And I appreciate the time and suggestions provided by the remainder of my committee, Larry Wasserman and Paul Fischbeck, and the inter-departmental ribbing about the lack of refreshments. Over the course of my graduate career, Jay Kadane and I enjoyed many interesting conversations, one of which led to my current (and much loved) job. Thanks to my fellow students Daniel Manrique and Han Liu for our many years of study groups and friendship.

When my grandmother was a young girl, her family lived in Johnstown, Pennsylvania, which was continually plagued by devastating floods. For a period of time, her father would walk every day to his job in the coal mine, put in a long day mining underground, walk home... and then spend his evenings single-handedly raising the family home onto stilts to try to protect it in advance of the next flood. In thanking my family, I of course owe much

to the patience of my husband and children and their forgiveness for all those weekends during which I went missing, and the long-time support and encouragement of my parents. But also I owe so much to everyone throughout my family tree who worked so hard at long, tedious, back-breaking work to slowly eke out just a little more for the next generation. I will not lose sight of how lucky I am to spend my days in my climate controlled office, seated in my comfortable chair, days spent working on problems that captivate me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the realm of high-dimensional statistics, regression and classification have received much attention, while research in density estimation has lagged behind. Yet there are compelling scientific questions which can only be addressed via density estimation using high-dimensional data; in particular, scientific questions which can be cast as probabilities of events. Consider the paths of North Atlantic tropical cyclones (TC), some of which are shown in Figure 1.1. How would one use this data to estimate the probability that a particular swath of coastal North Carolina will be hit by a TC in the next decade? Or how can one relate changes in TC paths over time to major climatic predictors such as sea surface temperature? If one models each track as a single draw from a probability distribution, estimation of this distribution allows one to answer such questions via integration or Monte Carlo methods. Important properties of a TC are highly dependent on its spatial positions going back to its genesis, and these terminal *and* intermediate positions are potentially related to large scale fluctuations in other properties of the climate system. Preserving the

entire track preserves the ability to discover any relationship between these fluctuations and, for example, landfall location.

1.1 Motivating problem

As defined in Jarvinen et al. (1984), tropical cyclones are "a nonfrontal low pressure system of synoptic scale[1] developing over tropical or subtropical waters[2] and having definite organized circulation[3]". The low-frequency, high-severity nature of tropical cyclones in the North Atlantic Ocean means that important and costly public policy, military, and business decisions are being made on the basis of relatively little historical data, and consequently any methodology that can extract more information from the data is very useful in advancing U.S. scientific, security, and economic interests. Much attention has been paid to hypotheses about the effect of various climatic predictors on TC occurrence frequency, TC landfall frequency, and TC intensity. However, few people have addressed the spatial variation of TCs, i.e. the TC tracks, and even fewer have exploited the relationship between climatic predictors and TC distribution. As Xie et al. (2005) state, in addition to the focus on yearly counts and intensity, "it would be of great benefit to society if the preferred paths of hurricanes could also be predicted in advance of the onset of hurricane season."

The statistical work in this thesis is motivated by a desire to improve the ability to answer questions about the preferred paths of hurricane tracks. Hurricane tracks are very

[1] A meteorological scale on the order of 1000 kilometers.
[2] Hence "tropical".
[3] Hence "cyclone".

high-dimensional objects — inherently infinite-dimensional, as they are curves, but represented in the standard database as a sequence of points representing a track's location at 6-hour intervals. The existing hurricane track estimation methods generally attempt to fit parametric models, but the complex, non-linear nature of the data requires an explosion of parameters to remain realistic. I demonstrate that a more nimble nonparametric approach to density estimation is more appropriate, one in which each track is represented as a single high-dimensional data point. This will lead to improved density estimates, and therefore improved public policy and business decisions.

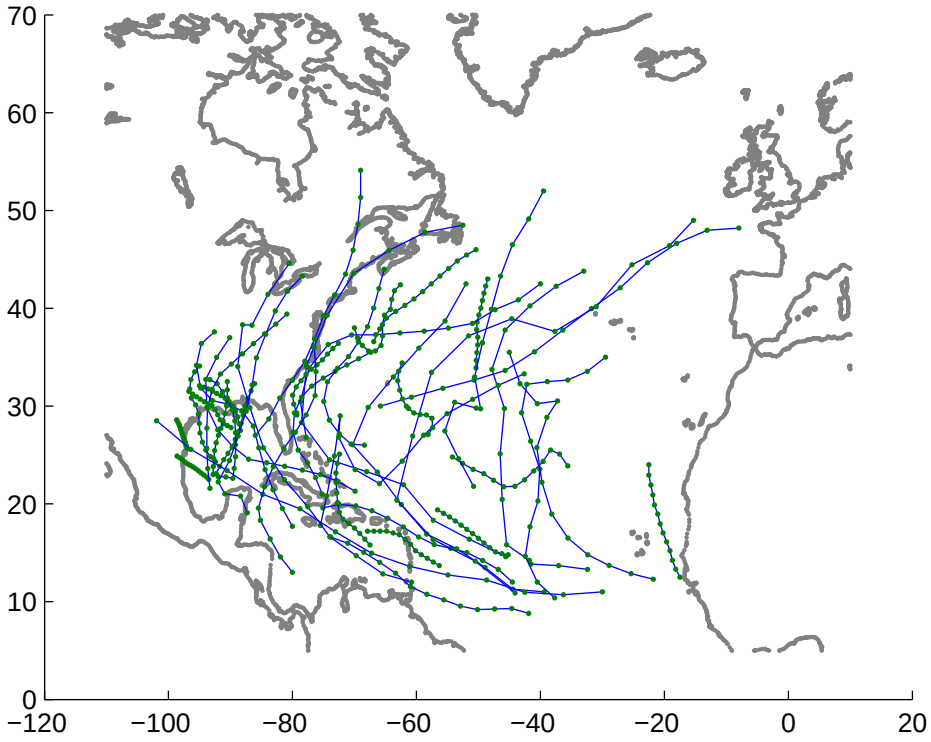


Figure 1.1: Forty randomly selected tropical cyclone tracks.

All attempts to perform high-dimensional density estimation (HDDE) will require an

element of dimensionality reduction to be feasible. Most existing methods, however, suffer from assumptions that are not appropriate for the application presented above, such as assuming that there is a suitable *linear* projection of the data into a lower-dimensional space. Furthermore, this application requires a method that not only provides density estimates for the observed data, but one in which any conceivable track can be mapped into the reduced space so as to obtain a density estimate. Without this property, one cannot use integration or Monte Carlo methods to ask arbitrary questions about TC behavior. Thus, there is a need for research on methods for nonparametric, nonlinear HDDE that involves dimensionality reduction and yet allows sampling from the original input space.

I present an approach which utilizes a *spectral connectivity analysis* (SCA) method (Lee and Wasserman, 2010) called diffusion maps. SCA transforms the data in a way that preserves context-dependent similarity. SCA and other eigenmap methods have been very successful for data parameterization (Coifman et al., 2005; Lafon and Lee, 2006; Belkin and Niyogi, 2003), regression (Richards et al., 2009), and clustering and classification (Ng et al., 2001; von Luxburg, 2007; Lafon and Lee, 2006; Belkin and Niyogi, 2004). In this thesis, I present an approach to density and point process estimation to better address key questions regarding TC behavior. However, this method is more broadly applicable to nonlinear, high-dimensional densities, particularly those for which there is an existing estimation method that incorporates significant amounts of relevant domain knowledge but which does not appropriately account for the dimensionality. For example, in the tropical cyclone realm, the knowledge could be of the land boundaries, theoretical upper limit of track speeds, etc. Many models exist to carefully account for this external domain

knowledge, yet still perform poorly as density estimates because of a lack of attention to the complexities of high-dimensional estimation (or because they are not stochastic models to begin with). The primary contribution of this thesis is a nonparametric, generalized method which essentially layers over such density estimates, via importance sampling, to keep the relevant domain knowledge while still accounting for dimensionality.

1.1.1 Data

As is standard in this field, I will rely on HURDAT, the "best track" database of North Atlantic tropical cyclones produced by NOAA, the National Oceanic and Atmospheric Administration (Jarvinen et al., 1984). Of the 608 TCs between 1950 and 2006, the longest TC was 1971's Hurricane Ginger, lasting 28 days; the year with the most storms was 2005, with 28; and the year with the fewest storms was 1983, with four. It is standard not to perform estimation on years before 1950; before then, only weather balloons, ship observations transmitted by telegraph, and organized naval reconnaissance were available as data collection methods, leading to undercounting and truncation in the database. The development of increasingly improved satellite measurements and data ocean buoys have improved records to the extent that data quality is not a primary concern in research using modern storms. Most tropical cyclone research that is not explicitly concerned with imputing the unobserved historical storms will work with a lower time bound between 1950 and 1975. In this thesis, I work with both 1950 and 1970 as lower bounds.

1.2 Existing climate work

The existing work modeling tropical cyclone tracks can be divided into two camps: "dynamical models" and "statistical models". Dynamical modelers are trying to create hurricane models from first physical principles, whereas the statistical modelers perform inference using historical tracks. NOAA's National Hurricane Center characterizes them as follows (National Hurricane Center, 2009):

> Dynamical models, also known as numerical models, are the most complex and use high-speed computers to solve the physical equations of motion governing the atmosphere. Statistical models, in contrast, do not explicitly consider the physics of the atmosphere but instead are based on historical relationships between storm behavior and storm-specific details such as location and date.

As a result of increasing computing power and improved understanding of the physics of the climate system, dynamical models are increasingly capable of resolving the behavior of TCs: they can generate suites of realistic tracks, in that they reside in the same low-dimensional space as real tracks. That does not mean, however, that they generate these tracks in the correct *distribution*, and hence they are not useful for estimating probabilities of interest. Or they generate tracks as a function of very specific initial conditions, and it is not clear how to generalize the results in order to make succinct claims about the past behavior of TCs on a much larger timescale, over which initial conditions are constantly changing.

One contribution of this dissertation is a statistical framework which can take advantage of dynamical models to inform the low-dimensional space in which tracks reside; this is the focus of Chapter 4. I will not discuss the details of any dynamical models, although they may be used in my work as described in Chapter 4. The details of statistical models, however, merit closer study so the comparison to our method can be clearly understood.

The majority of work in "statistical" track density estimation has adopted the following approach, working in two spatial dimensions: first estimate a genesis (origination) density over the region of interest (e.g. the North Atlantic); then estimate a series of Markovian densities of track propagation, usually corresponding to 6-hour steps in which the distribution of the next location is a function of only the previous location; finally couple this with a lysis (death) component so that the simulated hurricane eventually stops (Hall and Jewson, 2007; Rumpf et al., 2007; Emanuel et al., 2006; Vickery et al., 2000). For example, Vickery et al. (2000) uses the following model for changes in track speed c and direction θ of a TC from time i to $i+1$:

$$\begin{aligned} \text{Speed: } \triangle \ln \mathbf{c} &= a_1 + a_2\psi + a_3\lambda + a_4 \ln c_i + a_5\theta_i + \epsilon \\ \text{Direction: } \triangle\theta &= b_1 + b_2\psi + b_3\lambda + b_4 c_i + b_5\theta_i + b_6\theta_{i-1} + \delta \end{aligned}$$

where ψ and λ are latitude and longitude and ϵ and δ are error terms. In addition, to model spatial variability the parameters $a_1, a_2, \ldots, a_5, b_1, b_2, \ldots, b_6$ vary over *each box* in a $5^\circ \times 5^\circ$ grid over the Atlantic Ocean. Clearly, a primary drawback to this approach is the proliferation of parameters to estimate and models to validate.

1.2.1 TC tracks and climate factors

In Chapter 6, I study the effect that climatic factors have on TCs. The majority of work on these effects can be separated into two classes: correlating yearly counts (or some function thereof, such as percent change) with yearly averages of climatic factors (Elsner, 2006; Mann and Emanuel, 2006; Holland and Webster, 2007; Hall and Jewson, 2008; Oouchi et al., 2006; Saunders and Lea, 2008, 2005; Webster et al., 2005); and those which use the "modes" produced by methods such as empirical orthogonal function (EOF) analysis (Bell and Chelliah, 2006; Goldenberg et al., 2001; Kaplan et al., 1998; Xie et al., 2005).

For example, Hall and Jewson (2008) address the question of the effect of SST on landfall rates over fairly large regions of coastline; they use a rough-grained conditioning scheme which buckets years into "hot years" and "cold years" and search for differences in the two frequencies.

The modes of EOF are essentially the loading vectors resulting from Principal Component Analysis (PCA) (Scott, 1992) applied to a two-dimensional density. For example, Xie et al. (2005) extend beyond landfall considerations and use empirical orthogonal functions to correlate climatic predictors with a "hurricane track density function" (HTDF). However, HTDF is somewhat of a misnomer, as the object they construct is not a density over tracks but a density over the ocean: the magnitude of the HTDF at x corresponds to x's proximity to observed hurricane tracks.

It is important to note that these last two models — likelihood of landfall at different coastal regions or a density over the two dimensional ocean — can both be reduced to

an integral with respect to the density estimate for the full track space. Namely, one can think of them as the integrals over some set A, whether A is the set of all tracks that make landfall in North Carolina in a "hot year" or whether A is the set of all tracks that come close, however defined, to a particular point in the ocean. With a good density estimate, one can answer a whole variety of questions without creating specialized models for every avenue of investigation. This highlights an additional strength of my framework, which is that one need not create a separate, custom model for every question of interest or hypothesis.

1.3 Nonparametric high-dimensional density estimation

The alternative to the strong parametric assumptions of the previous models is to pursue *nonparametric density estimation.* Many nonparametric techniques for density estimation involve, in one form or another, estimating the density at x by smoothing over the proportion of data points in a neighborhood of x. But as discussed in the previous section, the data I am working with are high-dimensional, and therefore any attempt to naively apply standard nonparametric density estimation techniques will suffer from the "curse of dimensionality" (Scott, 1992; Donoho, 2000). As the neighborhoods grow in volume, the variance of our estimate decreases, but the bias increases; as the dimension grows linearly, the volumes must grow exponentially to contain the same number of data points. So one very quickly finds oneself ending up with either end up with a very unsmooth density estimate, or one in which the neighborhoods are so large that they obscure all local detail in the density.

Due to the curse, kernel density estimation (Scott, 1992), a very popular technique, would not be appropriate for this application, as there were only 608 storms between 1950 and 2006. As argued by Levina and Bickel (2004), "there is a consensus in the high-dimensional data analysis community that the only reason any methods work in very high dimensions is that, in fact, the data are not truly high-dimensional. Rather, they are embedded in a high-dimensional space, but can be efficiently summarized in a space of a much lower dimension, such as a nonlinear manifold." One can only proceed after making some reasonable assumptions about the data and providing a technique that can take advantage of those assumptions.

Linear methods, such as Principal Component Analysis (PCA) (Scott, 1992), simply project all data points onto a lower-dimensional hyperplane, and are hence not able to describe complex, nonlinear variations. More recent work in HDDE has assumed sparsity of the input data (Liu et al., 2007), in the sense that the complex variations in density are a function of only a few of the original coordinates used to represent a datum. This is not typical of the data I consider here.

In this dissertation, I will expand on the assumptions that are appropriate for the TC problem and introduce new density estimation methods – based on diffusion maps – to capitalize on this assumption. With the TC problem as a unifying thread, I will present two different density estimation methods, and one approach to conditional point processes. Chapter 2 will provide background on diffusions maps, a technique for dimensionality reduction that relates to a metric that quantifies the "connectivity" of the data set. Chapters 3 and Chapter 4 present two methods for high-dimensional density estimation using dif-

fusion maps. In Chapter 3, the data is projected into a lower dimensional space via the diffusion map and I perform standard, low-dimensional nonparametric density estimation methods there. In Chapter 4 the map serves as a basis adapted to an "instrumental distribution", such as the dynamical models of Section 1.2. This basis is used to construct a density for the observed data with respect to the instrumental distribution, which allows for integration of domain knowledge and more efficient Monte Carlo estimation. Chapter 5 presents the method of Chapter 4 applied to two simulation examples as well as the observed TC tracks from 1970 to 2006. The results show that the method improves on the conventional approaches to tropical cyclones, as reviewed in Section 1.2.

Lastly, Chapter 6 lays out some potential extensions to the work in this thesis. The chapter is mainly concerned with adding climatic predictors into the model, such as sea surface temperature. In that chapter, I show how our density estimation framework serves as a generalization to questions in the literature about TCs and climatic predictors, and I propose a method for conditional point process estimation using diffusion maps and presents some results.

Chapter 2

Diffusion Maps

In this chapter, we introduce diffusion maps, a technique for dimensionality reduction that relates to a metric that quantifies the "connectivity" of the data set and introduces a new coordinate system based on this metric (Coifman et al., 2005; Lafon and Lee, 2006; Lee and Wasserman, 2010; Freeman et al., 2009).

2.1 Introduction to diffusion maps

Assume that observed data $\Omega = \{X_1, \ldots, X_m\}$ are drawn from F with support $\mathcal{X} \subset \mathbb{R}^\ell$. We can consider Ω to be the vertices of a weighted graph $\Gamma = (\Omega, W)$, where the edge weights connecting $x, y \in \Omega$ are

$$k_\epsilon(x, y) = \frac{1}{(4\pi\epsilon)^{l/2}} \exp(-d^2(x, y)/2\epsilon), \tag{2.1}$$

where $d(x, y)$ is an application-specific locally relevant distance measure and ϵ controls the neighborhood size.

Suppose we now imagine a Markov random walk over this graph, where the probability of stepping directly from x to y is $p_1(x, y) = \frac{k_\epsilon(x,y)}{\sum_z k_\epsilon(x,z)}$. This probability will be very small unless x and y are similar to each other, i.e. $d(x, y)$ is small. The resulting one-step transition matrix $T = \{p_1(x, y)\}_{x,y\in\Omega}$ can be decomposed into a set of eigenvalues $1 = \lambda_0 \geq \lambda_1 \geq \ldots \lambda_{m-1}$ and left and right eigenvectors $\{\phi_i, 0 \leq i \leq m-1\}$ and $\{\psi_i, 0 \leq i \leq m-1\}$ where

$$\begin{aligned} \phi_j^T T &= \lambda_j \phi_j^T \\ T\psi_j &= \lambda_j \psi_j \end{aligned}$$

with the left eigenvectors normalized with respect to $1/\phi_0$ and the right eigenvectors with respect to ϕ_0, i.e. $||\phi_l||^2_{1/\phi_0} = \sum_x \frac{\phi_l^2(x)}{\phi_0(x)} = 1$ and $||\psi_l||^2_{\phi_0} = \sum_x \psi_l^2 \phi_0(x) = 1$. The term "diffusion map" arises from the idea that, having imagined a random walk on the graph, we can consider points x and y as close not just by, say, their Euclidean distance, but by the difference in their t-step conditional distributions $p_t(x, \cdot)$ and $p_t(y, \cdot)$. The *diffusion distance*

$$D_t(x, y) = ||p_t(x, \cdot) - p_t(y, \cdot)||^2_{1/\phi_0} = \sum_{z\in\Omega} \frac{(p_t(x, z) - p_t(y, z))^2}{\phi_0(z)} \tag{2.2}$$

has an advantage over other common metrics for distances between distributions in that,

in combination with the biorthogonal spectral decomposition of T^t:

$$p_t(x,y) = \sum_{j=0}^{m-1} \lambda_j^t \psi_j(x)\phi_j(y), \tag{2.3}$$

the diffusion distance can be re-expressed as

$$D_t(x,y) = \sum_{j=1}^{m-1} \lambda_j^{2t}(\psi_j(x) - \psi_j(y))^2 \tag{2.4}$$

and because of the decaying eigenvalues, one can truncate at a relatively small q yet still achieve an accurate approximation, as the latter $(m-q)$ terms do not contribute much to the sum:

$$D_t(x,y) \approx \sum_{j=1}^{q} \lambda_j^{2t}(\psi_j(x) - \psi_j(y))^2. \tag{2.5}$$

(More details on the selection of q and the relationship between ϵ and q are presented in Section 4.6.) Lastly, the *diffusion map*, defined as

$$\Psi_t : x \mapsto \begin{pmatrix} \lambda_1^t \psi_1(x) \\ \lambda_2^t \psi_2(x) \\ \vdots \\ \lambda_q^t \psi_q(x) \end{pmatrix}, \tag{2.6}$$

reduces to

$$D_t(x,y) \approx \sum_{j=1}^{q} \lambda_j^{2t}(\psi_j(x) - \psi_j(y))^2 = ||\Psi_t(x) - \Psi_t(y)||^2. \tag{2.7}$$

In other words, the mapping Ψ_t projects the data into $\mathbb{R}^q$ in such a way that the Euclidean distances in this diffusion space approximates the diffusion distance. Because the diffusion distance of Equation 2.2 sums over differences in conditional *distributions*, it is robust to noise; an $x, y \in \Omega$ that happen to be closely connected by a $z \in \Omega$ will still have a large diffusion distance if most paths between them are far relative to the average distance between pairs in Ω.

2.2 Connection to operators

In the previous section, we merely asserted ϵ's role as a smoothing parameter. Previously, we considered a Markov chain on the finite graph Γ; let us instead consider a Markov chain on the infinite state space $\mathcal{X}$, with the transition kernel

$$\Omega_\epsilon(x, A) = \mathbb{P}(x \mapsto A) = \frac{\int_A k_\epsilon(x,y)dF(y)}{\int k_\epsilon(x,y)dF(y)}. \tag{2.8}$$

This is a chain that moves from x to points y that are close to x, privileging those that reside in areas high in density $dF/d\mu$, where μ is Lebesgue measure. Let $p_\epsilon(x) = \int k_\epsilon(x,y)dF(y)$; the stationary distribution of the chain S_ϵ is

$$S_\epsilon(A) = \frac{\int_A p_\epsilon(x)dF(x)}{\int p_\epsilon(x)dF(x)} \tag{2.9}$$

and its density with respect to F is

$$s_\epsilon(x) = \frac{p_\epsilon(x)}{\int p_\epsilon(y)dF(y)}. \tag{2.10}$$

The diffusion operator A_ϵ is the continuous analog to the transition matrix T. Mapping a function $v \in \mathcal{L}_2$ to $A_\epsilon v$, the diffusion operator is defined as

$$A_\epsilon v(x) = \frac{\int k_\epsilon(x,y)v(y)dF(y)}{\int k_\epsilon(x,y)dF(y)}, \tag{2.11}$$

where its eigenfunctions $\{\psi_{\epsilon,0}, \psi_{\epsilon,1}, \ldots\}$ are orthonormal with respect to S_ϵ. In other words,

$$\int \psi^2_{\epsilon,i}(x)dS_\epsilon(x) = \int \psi^2_{\epsilon,i}(x)s_\epsilon(x)dF(x) = 1 \text{ and } \int \psi_{\epsilon,i}(x)\psi_{\epsilon,j}(x)dS_\epsilon(x) = 0. \tag{2.12}$$

Thus the eigenfunctions are an orthonormal basis with respect to S_ϵ, the stationary distribution for the Markov chain with smoothing parameter ϵ. Later, we will utilize the eigenvectors of T to estimate the eigenfunctions of A_ϵ and create an approximately orthonormal basis with respect to stationary distribution S_ϵ.

2.3 Nyström extension

The above construction creates diffusion map coordinates for only the members of Ω. To locate other objects in diffusion space, one must be able to extend the map to all members of $\mathcal{X}$. We rely on a technique known as the Nyström extension (Williams and Seeger,

2001). As the $\psi_{\epsilon,i}$s are eigenfunctions of A_ϵ, the equation

$$\lambda_{\epsilon,i}\psi_{\epsilon,i} = A_\epsilon \psi_{\epsilon,i}. \tag{2.13}$$

holds. If we substitute Equation 2.11 in Equation 2.13, we see that

$$\psi_{\epsilon,i} = \frac{A_\epsilon \psi_{\epsilon,i}}{\lambda_{\epsilon,i}} = \frac{\int k_\epsilon(x,y)\psi_{\epsilon,i}(y)dF(y)}{\lambda_{\epsilon,i}\int k_\epsilon(x,y)dF(y)}, \tag{2.14}$$

which justifies the estimator

$$\widehat{\psi}_{\epsilon,i}(x) = \frac{\sum_{j=1}^n k_\epsilon(x,X_j)\widehat{\psi}_{\epsilon,i}(X_j)}{\lambda_{\epsilon,i}\sum_{j=1}^n k_\epsilon(x,X_j)}. \tag{2.15}$$

Then the i^{th} basis function can be evaluated on any member of Ω^c as a weighted mean of the same basis function evaluated on the members of Ω. One can select an different smoothing parameter ϵ' to use in the kernel functions of Equation 2.15, for example:

$$\widehat{\psi}_{\epsilon,i}(x) = \frac{\sum_{j=1}^n k_{\epsilon'}(x,X_j)\widehat{\psi}_{\epsilon,i}(X_j)}{\lambda_{\epsilon,i}\sum_{j=1}^n k_{\epsilon'}(x,X_j)}. \tag{2.16}$$

So long as $\epsilon' = \epsilon$, members of Ω will remain unchanged under the Nyström extension.

2.4 Example

We provide some intuition about the information contained within the eigenvectors of Section 2.1 with an example on two-dimensional data. This example will help to make clear the potential of this dimension reduction technique for density estimation. Figure 2.1

displays a sample of 4000 from a distribution F; it is a compound distribution from which a mean is chosen along a spiral segment, then a normal deviate is selected perpendicular to the spiral.

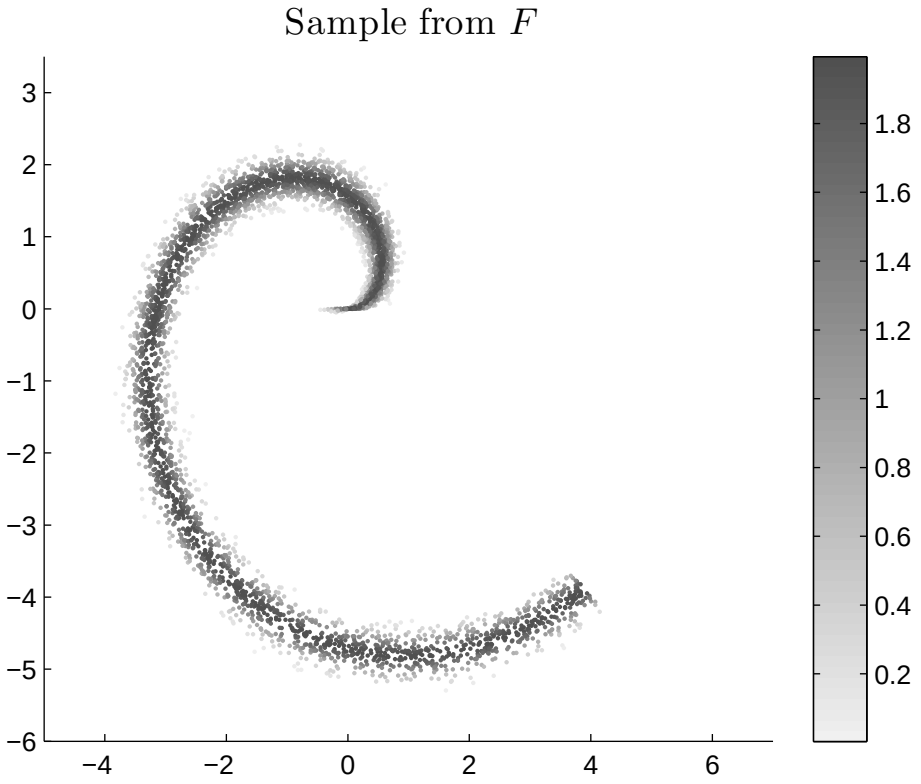


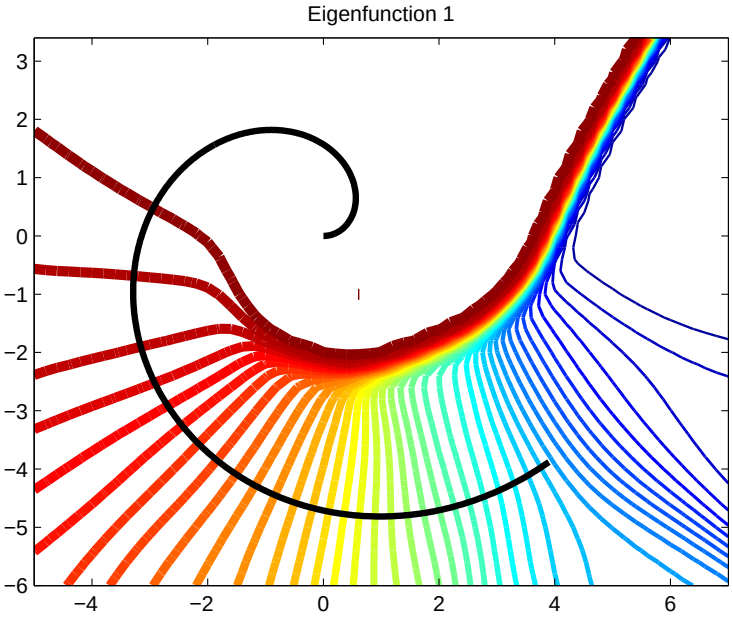
Figure 2.1: A sample of 4000 points from the distribution F.

The motivation of this particular example will be explained in Section 5.1, but taken at face value for now, the functions shown in Figures 2.2 and 2.3 illuminate what the process is doing. The zeroth eigenfunction is (necessarily) constant, and therefore omitted, but the first through fourth contours for the estimated eigenfunctions are instructive.

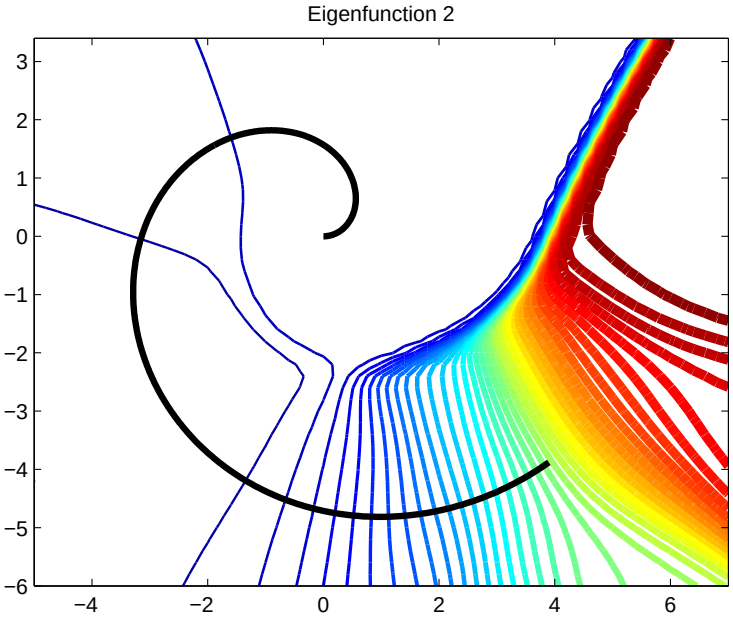
The first eigenfunction has contours *perpendicular* to the mean curve, and is essentially marking an observation's mean location. The second, third, and fourth eigenvectors all de-

scend steeply at the end of the mean spiral curve, which models the abrupt end to F there. However, each eigenfunction also has regions whose contours radiate somewhat *parallel* to the mean curve. The eigenfunctions essentially replicate the construction of the model — first, pick a location α on the mean curve, then select a point perpendicular to the curve at α, with higher probability given to points near to the curve. These eigenfunctions are estimated without any knowledge of F beyond the sample. Any nonparametric analysis relying on the functions — whether as a coordinate system, basis, or covariates in a regression — benefits from the fact that the functions are adapted to the problem at hand, because relative to, say, a basis chosen *a priori*, it reduces the number of functions needed to retain the same amount of information about the sample.

In this chapter, we have introduced diffusion maps and provided a low-dimensional example to motivate the eigenvectors that result from the maps. The eigenfuctions are well suited to use in HDDE because they are adapted to the problem at hand and robust to noise. The remaining chapters will present methods for using these vectors as a basis for density and point process estimation.

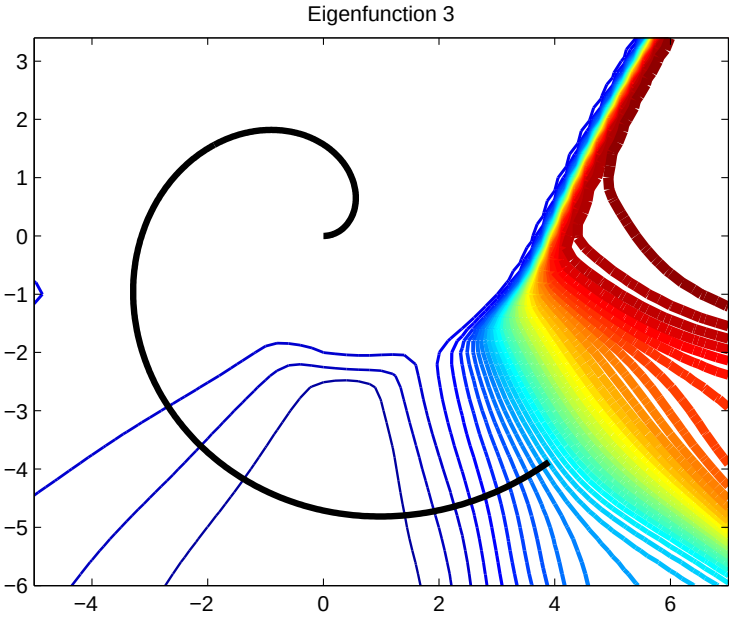


(a) $\widehat{\psi_1}$

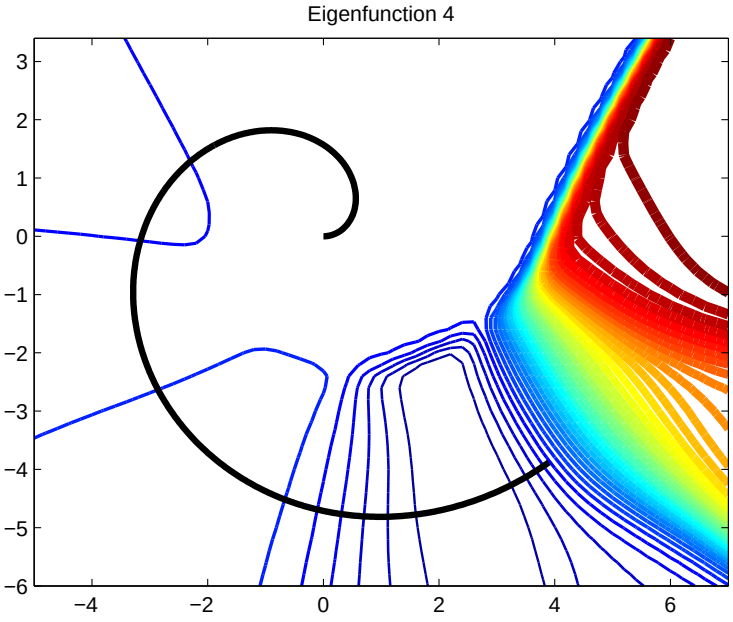


(b) $\widehat{\psi_2}$

Figure 2.2: **Eigenfunctions 1 and 2 for** F: The estimated eigenfunctions are smooth relative to F, and exhibit contours that mimic the data generation process itself — first, select a position on the curve, with higher density in f given to points toward the end of the curve; then, select a point perpendicular to the curve at that point. The thickness of each contour is proportional to its value.



(a) $\widehat{\psi_3}$



(b) $\widehat{\psi_4}$

Figure 2.3: **Eigenfunctions 3 and 4 for** F: The estimated eigenfunctions are smooth relative to F, and exhibit contours that mimic the data generation process itself — first, select a position on the curve, with higher density in f given to points toward the end of the curve; then, select a point perpendicular to the curve at that point. The thickness of each contour is proportional to its value.

Chapter 3

Density Estimation: The Coordinate System Approach

One established method for nonparametric, high dimensional density estimation is to project the data into a lower dimensional space and perform standard, low-dimensional nonparametric density estimation methods there. Scott (1992) discusses this approach applied to PCA and projection pursuit. This chapter explores this same concept by using diffusion maps as the dimensionality reduction technique, and then applying traditional kernel density estimation to produce a density in the reduced space.

A graphical overview of this method, as applied to TC tracks, is shown in Figures 3.1 — 3.3. While the details of the particular application are given in Section 3.3, the technique generally begins as follows: first, a diffusion map is constructed for the observed TC tracks, as shown in Figures 3.1(a) and 3.1(b). A kernel density estimate is then constructed in two-dimensional diffusion space, shown in Figure 3.2(a).

But how can one use the density of Figure 3.2(a) to ask questions about the physical world? What portions of the density in diffusion space correspond to, say, tracks which make landfall in Florida? Rather than trying to construct an interpretable physical map from all subsets of track space into all subsets of diffusion space, one can instead use Monte Carlo simulation. If one takes a very large draw from the density in diffusion space and maps each of those points back into track space, as shown in Figures 3.2(b) and 3.3(a), we can answer questions about the behavior of TCs by looking at the large sample mean.

More formally, let $\mathcal{X} \subset \mathbb{R}^\ell$ be *track space*, the space in which all possible TC tracks reside. Then for $\mathcal{X}$, subsets of track space $A \subset \mathcal{X}$ (for example, A could be the set of all tracks that never make landfall), a sample of observed tracks $\Omega \subset \mathcal{X}$, diffusion map $\Psi(\Omega)$, and an estimated density in diffusion space $\widehat{F}$, suppose that we want to estimate $\mathbb{P}(A)$. We create Φ, a large sample from $\widehat{F}$, invert the points of Φ back to track space:

$$\hat{\Omega} = \Psi^{-1}(\Phi) \tag{3.1}$$

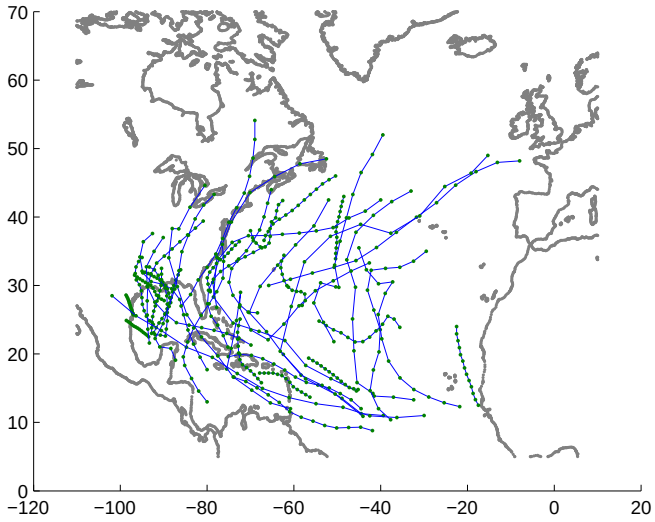
and estimate $\mathbb{P}(A)$ using Monte Carlo simulation:

$$\widehat{\mathbb{P}}(A) = \frac{1}{|\Phi|} \sum_{x \in \hat{\Omega}} I_{x \in A}. \tag{3.2}$$

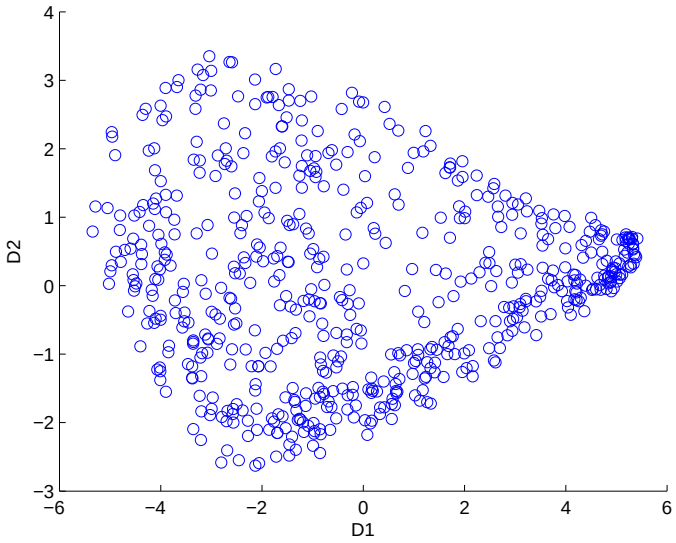
However, the inversion step can be quite challenging, as explained in the next section.

3.1 Details of method

The construction of the diffusion map was covered in detail in Chapter 2. The particular

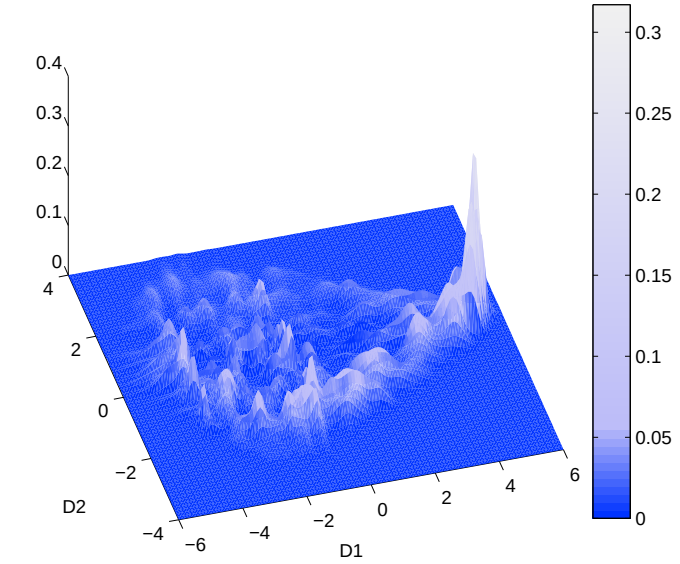


(a) A subset of the observed tracks.

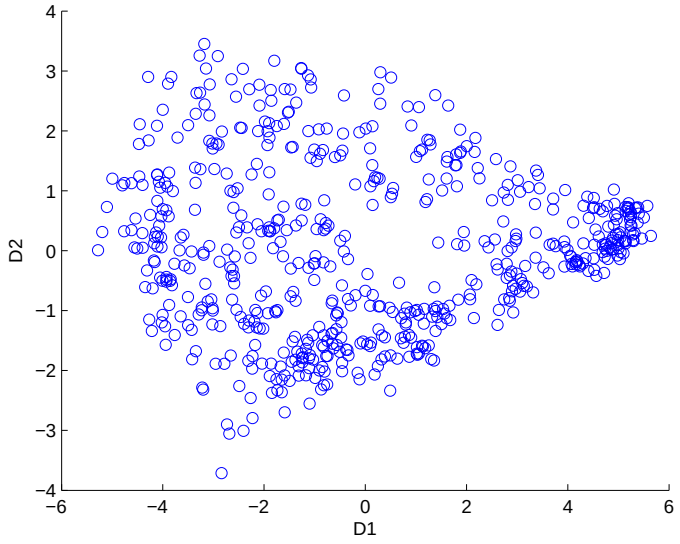


(b) The observed tracks in diffusion space.

Figure 3.1: **Part 1: An overview of the initial dimensionality-reduction approach to TC track simulation.** (a) shows 40 randomly selected tracks out of a total of 608 TCs observed between 1950 and 2006. (b) shows all 608 tracks mapped to a 2-dimensional diffusion space, with each point corresponding to a particular track in (a).

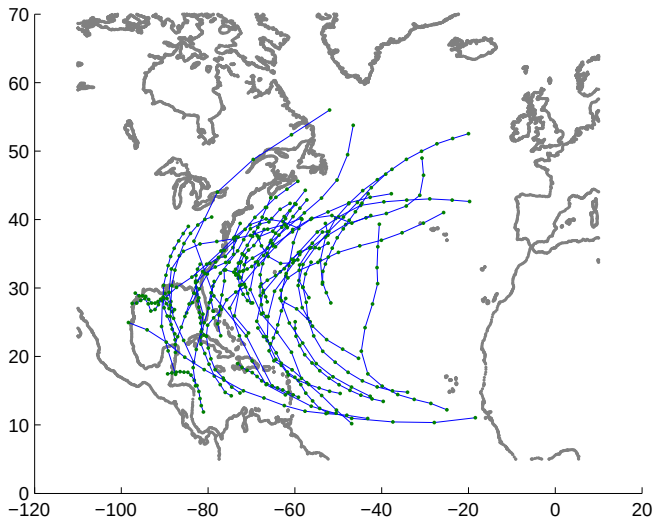


(a) A density over diffusion space.



(b) A random sample from the density.

Figure 3.2: **Part 2: An overview of the initial dimensionality-reduction approach to TC track simulation.** An estimated density for the diffusion space data of (b) is shown in (a), and a 608-element sample from that density is shown in (b). Each point in the sample can be interpreted as being associated with a new, as-yet-unobserved track. The sample here is the same size as the observed data for visual parity, but in practice one would use a very large sample from the density in diffusion space.



(a) A subset of the sample mapped back into track space.

Figure 3.3: **Part 3: An overview of the initial dimensionality-reduction approach to TC track simulation.** The sample is finally mapped back into track space; 40 randomly selected TCs of the sample are shown in (a).

Procedure 1 High-dimensional density estimation

Input: Ω, a high-dimensional data set of dimension ℓ and size m;
$d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, a locally-relevant distance measure.

Output: $\widehat{\Omega}$, a sample from the estimated density of Ω.

Dimensionality reduction:
    Construct Ψ, an q-dimensional diffusion map for Ω, d:
    Perform density estimation in diffusion space to form $\widehat{f}$.
    Generate Φ, a size-m random sample from $\widehat{f}$.
        Select model parameters ϵ, q, and σ_ζ via cross-validation using the method of Section 3.2.
Inverse mapping:
    Find $\widehat{\Omega} = \widehat{\Psi}^{-1}(\Phi)$, the pre-image of Φ in the ℓ-dimensional input space.
    Validate results via repeated simulation from $\widehat{f}$.

values of ϵ and q are chosen through cross-validation, where the particular loss function is introduces in Section 3.2. Once the data are mapped into the low-dimensional space, a range of nonparametric density estimators could be employed. We use k-nearest-neighbor kernel density estimation (rather than density estimation with a fixed kernel bandwidth), because of the tendency for data points to cluster near an apparent "boundary" in diffusion space; in particular, we use $k = \lfloor n \rfloor$ where $n = |\Omega|$, the size of the observed sample on which the diffusion map was built. When using this estimator, simulation is trivial (Silverman, 1986).

The limitation of this method is that inverting back to input space (i.e. track space) — transitioning from the step shown in Figure 3.2(b) to the one shown in Figure 3.3 — requires a search that is not necessarily convex and cannot be performed well in high-dimensional space. The section will first describe the inversion method, and then address its challenges.

We need a method for finding the pre-image of an arbitrary point in diffusion space. Using the Nyström extension of Section 2.3 there is a natural approach: the pre-image can be approximated as the track whose extension into diffusion space comes closest to the point that we wish to invert. In practice, however, designing a search mechanism that is both sufficiently exhaustive and computationally feasible is difficult. One solution is to restrict the pre-image to be a convex combination of observed data objects, assuming they are of the same dimension (or can be approximated as such in a meaningful way) (Kwok and Tsang, 2004; Mika et al., 1999). This is the approach we describe here.

Let ζ be the point in diffusion space for which we seek the pre-image. The Euclidean

distance from ζ to $\Psi(x)$ for each observed track $x \in \Omega$ is a natural measure of the similarity between $\Psi^{-1}(\zeta)$ and x; recall that, as shown in Equation 2.7, Euclidean distance in diffusion space approximates diffusion distance. Thus, we can construct weights as

$$w(\zeta, x) = \frac{\exp\left(-||\zeta - \Psi(x)||^2/\sigma_\zeta\right)}{\sum_{y\in\Omega} \exp\left(-||\zeta - \Psi(y)||^2/\sigma_\zeta\right)},$$

and then use these weights in constructing the convex combination:

$$\widehat{\Psi}^{-1}(\zeta) = \sum_{x\in\Omega} w(\zeta, x)\, x. \tag{3.3}$$

The inversion has thus been reduced to searching over a single parameter σ_ζ which controls the spread of the normal kernel used to determine the weights. This approach does require that each track be regularized to an equal number of points along its path (discussed in detail in Section 3.3), but that number need not be small in order for the construction of convex combinations to be feasible. Furthermore, note that in practice it will typically only be necessary to interpolate between tracks which are similar, i.e. if $w(\zeta, x)$ and $w(\zeta, y)$ are both large, then x and y will usually be alike. This is important because general convex combinations of tracks will not be physically-feasible tracks; on local scales, however, convex combinations are physically reasonable.

However, constructing inverses as convex combinations of observed tracks produces simulated tracks that are never more extreme than the most extreme observed track for whichever spatial measurement one might want to consider. For example, using this approach, no pre-image could be longer than the longest of the observed tracks or shorter

than the shortest of observed tracks. To overcome this shortcoming, in addition to searching over σ_ζ, we also allow the average to stretch up to 150% and shrink down to 75%. In other words, we do not search for merely the convex combination whose projection comes closest to η, but we treat the convex combination as a form which can be stretched (in both directions) by the optimal factor in $[.75, 1.5]$. Furthermore, we consider separately a shrink/stretch anchored at the origination point and the lysis point of the convex combination. Despite these extensions, one remaining shortcoming is that it is not possible to sample a track with more loops than any of the observed tracks, although it is possible for such a TC to occur. A set of 608 tracks simulated using our algorithm is shown in Figure 3.3(a).

Whether this restricted way of estimating pre-images is too narrow is just one of the many reasons one might want model validation. The entire procedure is summarized in Procedure 1, and the next section provides more details about validation. Finally, we present results in Section 3.3.

3.2 Validation

For choosing model parameters via cross-validation, we use an approach based on a test of the hypothesis that two high-dimensional samples — the observed data and simulated data — come from the same underlying distribution. Specifically, we produce a nonparametric high-dimensional verification technique that treats the particulars of the methodology as a black box. We assess whether a new sample can reasonably be said to come from the same distribution as the observed data, regardless of how the former was generated.

This is analogous to existing tools for one-dimensional analysis (Q-Q plots, the Wilcoxon rank-sum test, the two-sample Kolmogorov-Smirnov test). While there are multivariate extensions to some of these classic tests (Justel et al., 1997), these methods often struggle with extensions beyond two dimensions. We utilize a simple test statistic similar to that given in Hall and Tajvidi (2002), which allows for genuine high-dimensional comparisons, and also yields a visual assessment tool for helping to identify, and therefore possibly correct, the ways in which the simulation fails.

We make a connection between the choice of the local distance metric d and the validation of the density estimate; in practice, this connection can be used in motivating the choice of d. Formally, let μ_1 and μ_2 be two distributions over the input space and let $X_1, X_2, \ldots, X_n$ be i.i.d., distributed as μ_1, and let $Y_1, Y_2, \ldots, Y_n$ be i.i.d., distributed as μ_2. Define the quantity $\mathcal{L}_d(\mu_1, \mu_2)$ to be the proportion of the values

$$(X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n) \tag{3.4}$$

whose nearest neighbor (as measured by d) is from the same sample. Let $\mathcal{R}_d(\mu_1, \mu_2) = \mathbb{E}\left(\mathcal{L}_d(\mu_1, \mu_2)\right)$. We define a density estimator to be *consistent with respect to local distance metric* d if

$$\lim_{n\to\infty} \mathcal{R}_d(\widehat{\mu}_n, \mu_X) = 0.5, \tag{3.5}$$

where $\widehat{\mu}_n$ is the estimated distribution, and μ_X is the true distribution. Heuristically, if the two distributions $\widehat{\mu}_n$ and μ_X are the same, then the nearest neighbor of any sample value is equally likely to be from either of the two samples.

A simulated test

In practice, how can one use the motivation behind the more formal notion of consistency with respect to the local distance metric to produce a test of our sampling mechanism? Noting that all samples generated by the algorithm will be from the same distribution, we can used a simulation-based approach:

1. For some large number k, generate k pairs of samples of size n using the algorithm.
2. For the i^{th} pair, calculate and record $\mathcal{L}_d$.
3. Generate one last sample of size n using the algorithm and pair it with the observed tracks; calculate $\ell^* = \mathcal{L}_d$ for these values.
4. Evaluate where ℓ^* falls in the distribution of the k proportions; reject the hypothesis that the observed tracks come from the estimated density if it is too far in the tails.

This test can be adapted to any sampling mechanism, not just the one presented in this chapter.

Visual assessment

In a Q-Q plot one can often immediately diagnose the nature of dissimilar samples (for example, one sample having heavier tails than another). However, as emphasized in Hall and Heckman (2002), it is harder to craft visualization methods for high-dimensional data, as they tend to be co-dimensional with the data. But if one of the samples is created via a method that involves dimensionality reduction, this can be used in conjunction with the local distance metric to provide a quick visual gauge of the region of dissimilarity.

We plot both the original data and the sampled data in diffusion space, distinguishing the points not based on which sample they came from but by whether their nearest neighbor is of the same or different sample. One might be able to visually identify a region in the diffusion map which is too saturated with data points who have within-sample nearest neighbors. Of course, this will not provide as easy an answer as "different thickness of tails" or other causes of one-dimensional dissimilarity, but by inspecting the data points in the saturated region in their higher-dimensional representation, it can provide one with tools for generating hypotheses on why the simulation is insufficient which a single test statistic would not be able to do.

3.3 Results

To apply the methods of this chapter to the TC data, we first need to select a distance metric d. For this application, we are primarily interested in spatial variation alone, i.e. Did two TCs take similar paths?, rather than spatio-temporal variation, i.e. Did two TCs take similar paths at the same speed? The distance metric that we selected involves first regularizing each track so that each track is made of up $j = 12$ equi-distant segments. An example of this is shown in Figure 3.4. $d(x, y)$ is then the sum of the thirteen Great Circle distances (in kilometers) (Sinnott, 1984) between corresponding regularized points on the pair of tracks.

After selecting a distance metric, one needs to select an ϵ for the diffusion map (we choose $t = 1$ a priori) and an q for the dimension of the reduced space. Note that there is a tradeoff between the dimensionality reduction — in which a larger q improves the results

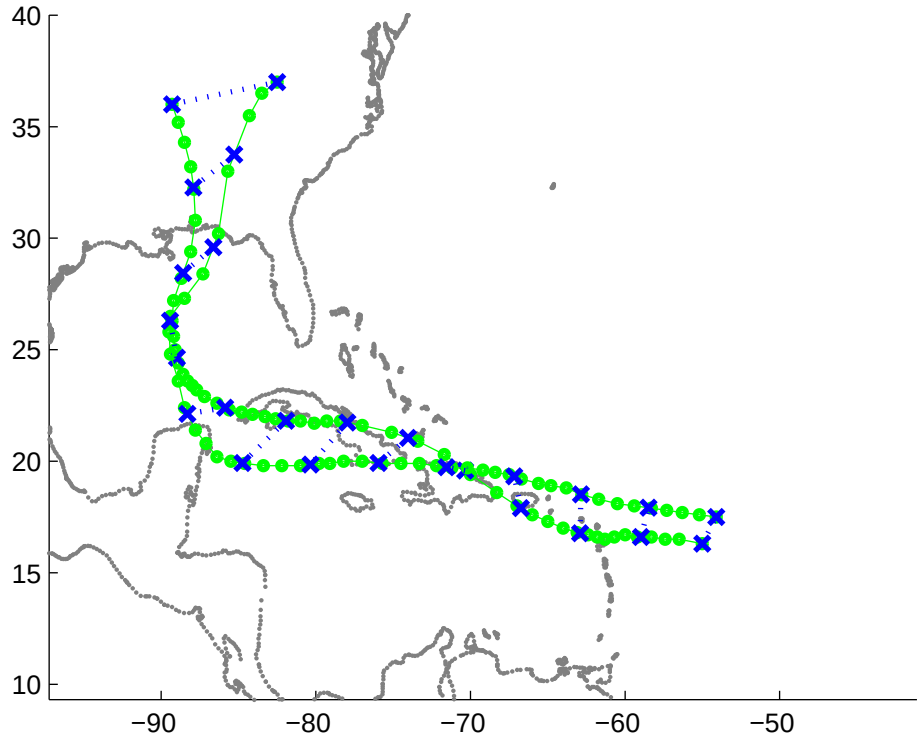


Figure 3.4: Regularization of tracks. Two regularized tracks are shown above: the original 6-hour segments are marked by circles, and the regularized segments are marked by $\times$. The application-specific distance measure is the sum of the Great Circle distances between 13 corresponding pairs of points, shown as dashed lines.

by retaining more information — and the density estimation — in which a larger q makes density estimation more difficult. We use the test of the previous section to choose not just ϵ but q and σ_ζ, too. We perform a search over $(\epsilon, q, \sigma_{zeta})$ and select the triplet whose within sample nearest neighbor proportion comes closest to 0.5. For this application we considered $q \in \{2, 3, 4\}$, as kernel density estimation in greater than four dimensions is not feasible with only 608 observations. When this test was applied to the observed tracks and the sample shown in Figure 3.3(a), the optimal pair was $q = 3$ and $\epsilon = 5217$.

In applying the test of the previous section, there were 689 within-sample nearest neighbors, for a proportion of $\frac{689}{2 \cdot 608} = 0.56$. For $k = 1000$, there were 31 pairs whose within-sample nearest neighbor proportion was higher, despite being from the same distribution, so the data has an estimated p-value of 0.031. This indicates that there is room for improvement in the steps of our algorithm, but the simulated data are fairly similar to the observed data. Given that we used only $q = 3$, this is quite encouraging.

Chapter 4

Density Estimation: The Basis Approach

The density estimation procedure of Chapter 3 performs reasonably well for TC tracks, but it suffers from the limitation that: a) it uses an inversion process which does not easily allow for simulated points that are more extreme than the observations, and b) despite this restriction, the inversion process can still be very cumbersome. The density estimation method of this chapter is best suited to fields such as climate science in which there has been a lot of work towards characterizing the space of realistic observations, leading to the capability to sample from this space, although not necessarily in the right distribution. As in the previous chapter, we use Monte Carlo methods — specifically, importance sampling — in conjunction with the eigenvectors of the diffusion map, in order to estimate quantities of interest.

4.1 High-dimensional importance sampling

Monte Carlo (MC) integration is a fundamental tool in the calculation of high-dimensional integrals. The method uses sample means of appropriately sampled random variables to estimate integrals that can be expressed as an expectation with respect to some target distribution. Importance sampling (IS) is a specialized form of MC integration designed to reduce the variance of the MC estimator; the random variables are rescaled by an instrumental density that is (ideally) chosen to specifically to minimize the variance of a particular integral (Liu, 2008).

Almost all applications assume that the target density is known and the only challenge is to find an instrumental distribution that not only reduces the variance of the estimator, but can be easily sampled from (Zhang, 1996; Berliner, 2001; Bengtsson et al., 2003). It is understood that this will be especially challenging in a high-dimensional setting (Bengtsson et al., 2008). However, the integral of interest might be with respect to an *estimated* density for an observed data set. With traditional importance sampling, this would require three steps: estimation of the high-dimensional density, selection of the appropriate instrumental distribution, and finally estimation of the integral with MC integration. For many high-dimensional applications, there are very few applicable density estimation methods (Buchman et al., 2011), making the first step difficult.

We propose a re-ordering of the estimation steps for this setting: first select an appropriate instrumental distribution, then proceed to estimation of the density, and finally estimate the integral. In many applications there exists a plausible simulation model – one which is known to refrain from generating unreasonable samples and to generate rea-

sonable samples, although not in the same proportions as the underlying data distribution that is being estimated. The key to our approach is to perform not the standard density estimation with respect to the Lebesgue measure, but instead derive a density estimate for the data distribution with respect to the distribution of the plausible simulation model. The simulation model serves as the instrumental distribution. *Rather than fixing a high-dimensional density estimate and constructing an instrumental distribution to minimize variance, we fix an instrumental distribution and construct a density estimate that reduces variance via dimensionality reduction.*

Specifically, for data distribution F and instrumental distribution G, with respective densities $f(x) = dF/d\mu$ and $g(x) = dG/d\mu$ for Lebesgue measure μ, traditional importance sampling relies on the ratio of the two densities:

$$\mathbb{E}_F(h(X)) = \int h(x)f(x)\mu(dx) = \int \frac{h(x)f(x)}{g(x)}g(x)\mu(dx) = \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} h(X_i)\frac{f(X_i)}{g(X_i)} \text{ a.s.} \tag{4.1}$$

where $X_1, \ldots, X_m$ are drawn from G. In the density estimation context, $f(x)$ would be an estimate of a high-dimensional density. Having somehow arrived at a density estimate, traditional high-dimensional importance sampling would proceed by constructing an instrumental distribution that is tailored to "target function" $h(x)$, the function whose expected value is desired. For the IS estimator of Equation 4.1, it can be shown that the optimal instrumental distribution is $g(x) = |h(x)|f(x)/\int |h(x)|f(x)dx$ (Zhang, 1996). This, of course, relies on the very integral that we are trying to calculate, and must be ap-

proximated. This can be done parametrically (Oh and Berger, 1993) or otherwise (Zhang, 1996), but would result in a ratio of two high-dimensional density estimates, where $g(x)$ must be re-derived for every $h(x)$.

In our method, rather than estimating $f(x)$, we estimate $f(x)/g(x) = dF/dG$ directly for an appropriately chosen simulation model G. This density does not have to vary with h and avoids the need to estimate two high-dimensional densities. The approach relies on diffusion maps to create a lower-dimensional, over-smoothed estimate of G. In combination with a kernel density estimate, this over-smoothed estimate can be used to construct an orthogonal series density estimate of the data distribution with respect to the instrumental distribution. We call this procedure *adaptive basis density estimation* (ABDE), as the basis is adapted to the instrumental distribution.

This method is especially applicable to the problem of "simulation calibration", i.e. parameter estimation for models proposed "without explicitly modeling the relationship between truth and the simulator" (Campbell, 2006). An advantage of our method is that is does not require the density function of the instrumental distribution, so simulation models without explicit probability models attached are acceptable; we only require that the instrumental distribution can be sampled from cheaply. In many model calibration efforts, it is not reasonable to assume that the simulator can be run quickly, however – hence the widespread interest in "emulators", which attempt to predict the response of the simulation with untried parameter values as a result of responses to parameters that were tried (O'Hagan, 2006; Annan and Hargreaves, 2007). There has been recent work in emulation using faster, simpler versions of complex models. These could serve as

instrumental distribution models in our framework, and runs of the complex simulation model would serve as the observed data.

4.2 Overview of estimation method and intuition behind it

The idea behind our method is as follows. The goal is to perform density estimation for data from high-dimensional distribution F. Suppose there exists a distribution G which puts most of its mass in the same region as the ℓ-dimensional *data distribution* F, although not in the same proportions. There need not be a known density for G — to return to the examples of Section 1.2, it might be a dynamical model. If one could sample inexpensively from this distribution *and* if one could estimate dF/dG, then most questions that could be answered with the more traditional $dF/d\mu$, where μ is the Lebesgue distribution, could be answered via rejection and importance sampling.

Our construction of dF/dG is built up as follow:

$$\boxed{f = \frac{dF}{dG} = \frac{dF}{dS_\epsilon} \cdot \frac{dS_\epsilon}{dG} = p^* \cdot s_\epsilon.} \tag{4.2}$$

Recall from Section 2.2 that S_ϵ is the stationary distribution of the imagined random walk of a large sample from G, $p^* = \frac{dF}{dS_\epsilon}$ and, as before, $s_\epsilon = \frac{dS_\epsilon}{dG}$. The first factor of the right-hand side of Equation 4.2 is estimated using orthonormal series estimation, where the series in question is the estimated eigenfunctions of A_ϵ (also introduces in Section 2.2). The second factor is estimated using a kernel density estimator; the key here is that, although we are attempting a high-dimensional density estimate using KDE, because we can sample

from G cheaply, we can sample as many data points as we would like and improve the accuracy of the density estimate.

4.2.1 Our contribution

Certainly the idea of orthogonal series density estimation is not new (Crain, 1973; Girolami, 2002; Efromovich, 1999; Kronmal and Tarter, 1968; Diggle and Hall, 1986; Bunea et al., 2007). Most methods, however, attempt to start with a very large basis and somehow pare down the number of basis functions used in the estimation. This does not allow one to integrate existing knowledge in the applied field, and customizing the basis to the problem at hand might be more successful.

Girolami (2002) employs very similar reasoning, although he is working with the un-row-normalized graph, i.e. the edge weight between x and y is $k_\epsilon(x, y)$, not $p_1(x, y)$. This results in some major differences. Firstly, the density he is estimating is with respect to the Lebesgue measure; as discussed above, we feel that this is less likely to be successful in high dimensions. Furthermore, he is assuming that the smoothing parameter ϵ is given, whereas we feel that a careful selection of ϵ is a crucial component of the work. Lastly, again there is no natural way to take advantage of the subject matter expertise inherent in existing models and the realistic tracks that they generate.

4.3 Details of method

We start with an m-member sample Ω_S, the members of which are assumed to be an iid sample from F. We begin by generating Ω_I, a large sample of size m from the instrumental

distribution G. We will work more directly with the stationary distribution S_ϵ than in Chapter 3.

4.4 Estimating s_ϵ

How do we estimate $s_\epsilon = dS_\epsilon/dG$? Recall from Equation 2.10 that

$$s_\epsilon(x) = \frac{p_\epsilon(x)}{\int p_\epsilon(y) dG(y)} \tag{4.3}$$

and

$$p_\epsilon(x) = \int k_\epsilon(x, y) dG(y). \tag{4.4}$$

Thus it is obvious that we can estimate p_ϵ as

$$\widehat{p}_\epsilon(x) = \frac{1}{m} \sum_{i=1}^{m} k_\epsilon(x, X_i). \tag{4.5}$$

To estimate $\int p_\epsilon(y) dG(y)$, the normalizing constant for s_ϵ, we note that

$$\int p_\epsilon(y) dG(y) \approx \int \left(\frac{1}{m} \sum_{i=1}^{m} k_\epsilon(y, X_i) \right) dG(y). \tag{4.6}$$

But we cannot merely take another sample mean of a function of the instrumental sample, as the random variables are no longer independent (in other words, $k_\epsilon(x, X_h)$ and $k_\epsilon(x, X_j)$ are independent for $h \neq j$, but $\frac{1}{q}\sum_{i=1}^{q} k_\epsilon(X_h, X_i)$ and $\frac{1}{q}\sum_{i=1}^{q} k_\epsilon(X_j, X_i)$ are not). This can be easily dealt with by generating another sample from G, call it S', of size m', and

taking the second sample mean over this new data set:

$$\int p_\epsilon(y)dG(y) \approx \frac{1}{m'}\sum_{j=1}^{m'}\frac{1}{q}\sum_{i=1}^{q}k_\epsilon(X'_j, X_i). \tag{4.7}$$

Thus

$$\widehat{s}_\epsilon(x) = \frac{\frac{1}{q}\sum_{i=1}^{q}k_\epsilon(x, X_i)}{\frac{1}{m'}\sum_{j=1}^{m'}\frac{1}{q}\sum_{i=1}^{q}k_\epsilon(X'_j, X_i)} \tag{4.8}$$

4.5 Estimating p^*

In Section 2.3, we described how to estimate $\widehat{\psi}_{\epsilon,i}$, which approximates an orthonormal basis with respect to S_ϵ. Using orthogonal series density estimation, we now estimate p^* using that basis:

$$p^*(x) = \sum_{i=0}^{\infty}\alpha_i\psi_i(x). \tag{4.9}$$

Having chosen a basis which is orthonormal with respect to S_ϵ, we can easily find an unbiased estimator for the α_js:

$$\begin{aligned}
\mathbb{E}_F(\psi_j(Z)) &= \int \psi_j(z)dF(z) \\
&= \int \psi_j(z)f(z)dG(z) \\
&= \int \psi_j(z)p^*(z)s_\epsilon(z)dG(z) \\
&= \int \psi_j(z)\left(\sum_{i=0}^{\infty}\alpha_i\psi_i(z)\right)s_\epsilon(z)dG(z) \\
&= \sum_{i=0}^{\infty}\int \alpha_i\psi_j(z)\psi_i(z)s_\epsilon(z)dG(z) \\
&= \alpha_j
\end{aligned}$$

suggesting the unbiased estimator

$$\widehat{\alpha}_j = \frac{1}{m}\sum_{i=1}^{m}\psi_j(X_i). \tag{4.10}$$

In practice, we of course do not use an infinite basis as in Equation 4.9 — selection of the q-term truncation point is discussed in the next section, and will be a function of sample size, the geometry of F, and the smoothness of F relative to G. But the distinction between this and most other methods is that the first few bases are not selected merely because they are smoothest relative to Lebesgue measure, but because they are smooth relative to (an aspect of) the distribution for the data of interest.

So, our final estimator for p^* is

$$\widehat{p}^*(x) = \sum_{i=0}^{q}\widehat{\alpha}_i\widehat{\psi}_{\epsilon,i}(x), \tag{4.11}$$

where $\widehat{\alpha}_i$ was defined in Equation 4.10 and $\widehat{\psi}_{\epsilon,i}(x)$ was defined in Equation 2.15.

4.6 Selection of ϵ and q

We have, up until this point, treated ϵ and q as fixed in ABDE, but unlike related work which side-steps the issue of bandwidth selection (Girolami, 2002), we want the selection of this value to be data-driven. Both parameters control smoothing/over-fitting — *increasing* ϵ leads to greater smoothing, whereas *decreasing* q increases bias but reduces variance.

There is work devoted to plug-in methods for selecting q in the orthogonal series estimation literature, but it generally assumes a fixed basis (Kronmal and Tarter, 1968; Diggle and Hall, 1986; Hart, 1985; Efromovich, 1999). The presence of ϵ means that we are not only determining how many basis functions to use but *which basis* to use, as each ϵ will produce a different one. To see this, merely refer back to Equation 2.15 and consider the effect that ϵ has on the estimators.

Whether one assumes a fixed or adaptive basis, most methods for choosing an optimal stopping rule, i.e. choosing q, centers around minimizing mean integrated squared error (MISE):

$$\begin{aligned} MISE(q,\epsilon) &= \mathbb{E}\left(\int (\widehat{f}_{q,\epsilon}(y) - f(y))^2 dG(y)\right) \\ &= \mathbb{E}\left(\int \widehat{f}^2_{q,\epsilon}(y) dG(y) - 2\int \widehat{f}_{q,\epsilon}(y) f(y) dG(y) + C\right) \\ &= \mathbb{E}\left(\int \widehat{f}^2_{q,\epsilon}(y) dG(y) - 2\int \widehat{f}_{q,\epsilon}(y) dF(y)\right) + c \\ &\equiv J(q,\epsilon). \end{aligned}$$

As in conventional MISE calculations, the second integral $\mathbb{E}\left(\int \widehat{f}(y)dF(y)\right)$ is estimated unbiasedly as a sample mean using leave-one-out cross validation. Unlike in conventional MISE calculations, the measure that the squared error is being integrated with respect to here is a probability distribution, in which case the random variable of the first expectation is an expected value, too, with respect to the instrumental distribution and can also be estimated by a sample mean: $\mathbb{E}\left(\int \widehat{f}^2(y)dG(y)\right) = \mathbb{E}\left(\mathbb{E}(\widehat{f}^2(Y)|Y)\right) = \mathbb{E}\left(\widehat{f}^2(Y)\right)$. In other words, using Ω_Z, a sample from G of size m_{CV},

$$\widehat{J}(q,\epsilon) = \frac{1}{m_{CV}}\sum_{i=1}^{m_{CV}} \widehat{f}^2_{q,\epsilon}(Z_i) - \frac{2}{m}\sum_{i=1}^{m} \widehat{f}_{(i),q,\epsilon}(X_i)$$

where $\widehat{f}_{(i),q,\epsilon}$ denotes the density estimate constructed using all data points but the i^{th} one. In this thesis, ϵ and q for ABDE are selected by minimizing $J(q,\epsilon)$ by search.

4.7 Importance and rejection sampling

In order to make use of our estimate $\widehat{f}$, we need to perform either importance or rejection sampling. Recall that for function $h(x)$, basic importance sampling (Robert and Casella, 2004) states that for $X_i \sim G$, $E_F(h(x)) = \int h(x)\frac{dF(x)}{dG(x)}dG(x) \approx \frac{1}{m}\sum_{i=1}^{m} h(X_i)f(X_i) \approx \frac{1}{m}\sum_{i=1}^{m} h(X_i)\widehat{f}(X_i)$.

For example, in the TC context, $h(x)$ might be an indicator function which is 1 if a track makes landfall in some region. Again, because we can sample cheaply from G, estimating the expected value of interest can be done with arbitrary precision.

Rejection sampling poses a bit more of a challenge in that we need a bound c such that $\forall x, \widehat{f}(x) \leq c$. We can bound $\widehat{f}$ by noting that because the estimated eigenfunctions $\mathcal{X}$ are a weighted sum of a finite eigenvector, scaled by the eigenvalues, then for $c_0 = \sup_x \widehat{f}(x)$,

$$c_0 \leq \left(\sum_{i=0}^{q} \frac{|\widehat{\alpha}_i|}{\lambda_{\epsilon,i}} max_{j:\mathtt{sgn}(\psi_i(X_j))=\mathtt{sgn}(\widehat{\alpha}_i)} |\psi_i(X_j)| \right) = c. \tag{4.12}$$

The next chapter provides a demonstration of ABDE on two simulation examples and on the TC data.

Chapter 5

Applications of ABDE

In this chapter, we present results of the ABDE method. First are two simulation examples — one low-dimensional, and one in the TC context – followed by an application of ABDE to the TC data.

5.1 Two dimensional example

Recall the example of Section 2.4, in which we presented some eigenvectors associated with the data distribution shown in Figure 2.1. We present an example of the power of the method of Section 4 with an example on two-dimensional data. The low dimension makes visualization possible, but I chose the size of the observed data set to be small enough to mimic the high-dimensional situation. The 20 points for which we want to perform density estimation are shown in Figure 5.1. Their underlying distribution F is a compound model — the mean is selected uniformly along a spiral segment, and the data point is chosen as a normal variable perpendicular to the spiral at its mean. We also assume that we

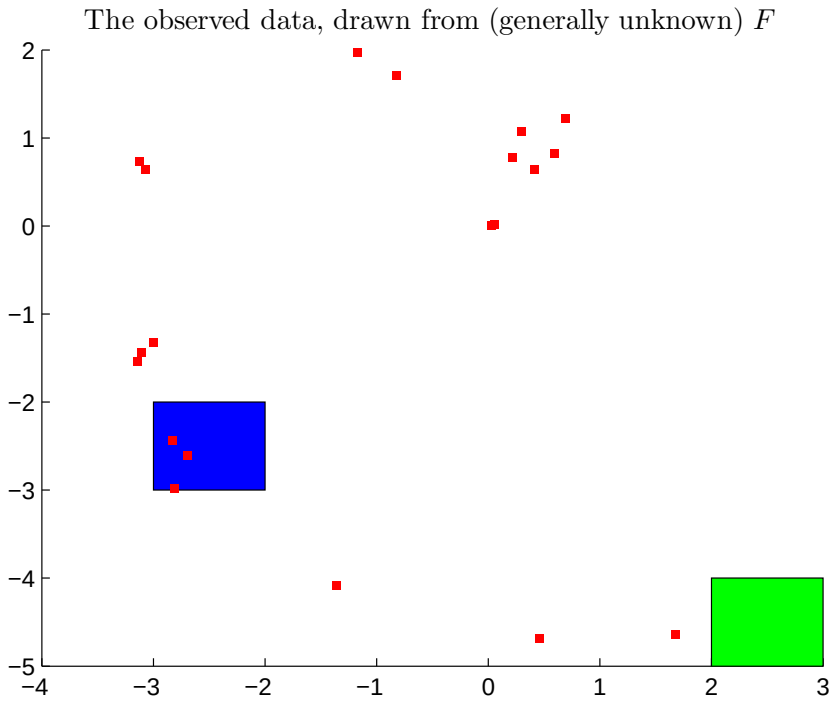


Figure 5.1: The 20 observed data points are shown as the small red squares. The green (light) and blue (dark) rectangles correspond to sets on which different measures can be compared in Table 5.1.

	F	G	$\widehat{f}$	KDE
$\mathbb{P}([2,3] \times [-5,-4])$	0.025	0.0042	0.024	0.0056
$\mathbb{P}([-3,-2] \times [-3,-2])$	0.033	0.015	0.038	.047

Table 5.1: The table compares the true measure of a set under both F and G to the estimates using adaptive basis estimation and kernel density estimation.

have instrumental distribution G, which takes the same general compound form but uses a different, longer spiral for the mean, does not select uniformly from that mean, and has a larger variance for the normal variable. Two large samples from each distribution are plotted in Figure 5.2, shaded by their true density with respect to μ.

Given $n = 20$ and the two-dimensional data, we selected $m = 4000$ and $d(x, y) = ||x - y||$, and using the methodology of Section 4.6, $\epsilon = .35$ and $q = 4$ were chosen. Figure 5.3 shows the true versus estimated $\frac{dF}{dG}$, and we see that the method performs well, even in regions where dF is greater than 5 times dG. Table 5.1 compares the probability of being a specified region for four distributions: the true F; F as estimated by $\widehat{f}$; F as estimated by kernel density estimation; and G. (The two regions can be seen relative to the observed data in Figure 5.1; one region does not contain any data points.) The last one is provided to make clear that the method is not swamped by G; in fact, we can see from the first line of the table that in the first region (denoted by the green rectangle in Figure 5.1) where the true probability under G is an order of magnitude less than that of F, the estimate is quite close. We see that in both regions, adaptive basis ODE outperforms KDE, where the latter's bandwidth was selected via leave-one-out cross validation.

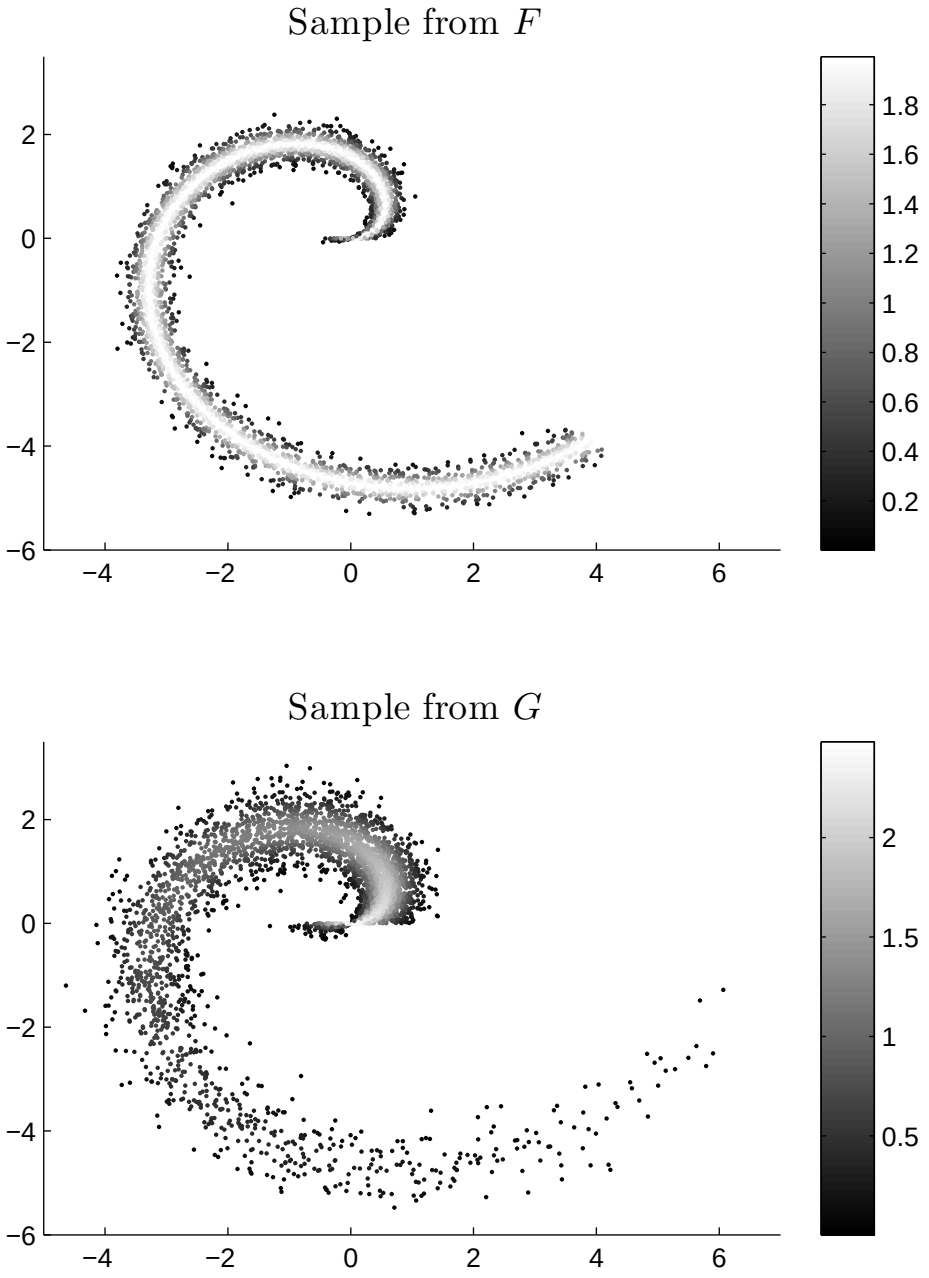


Figure 5.2: A sample of 4000 points from the distributions F and G for the two-dimensional example.

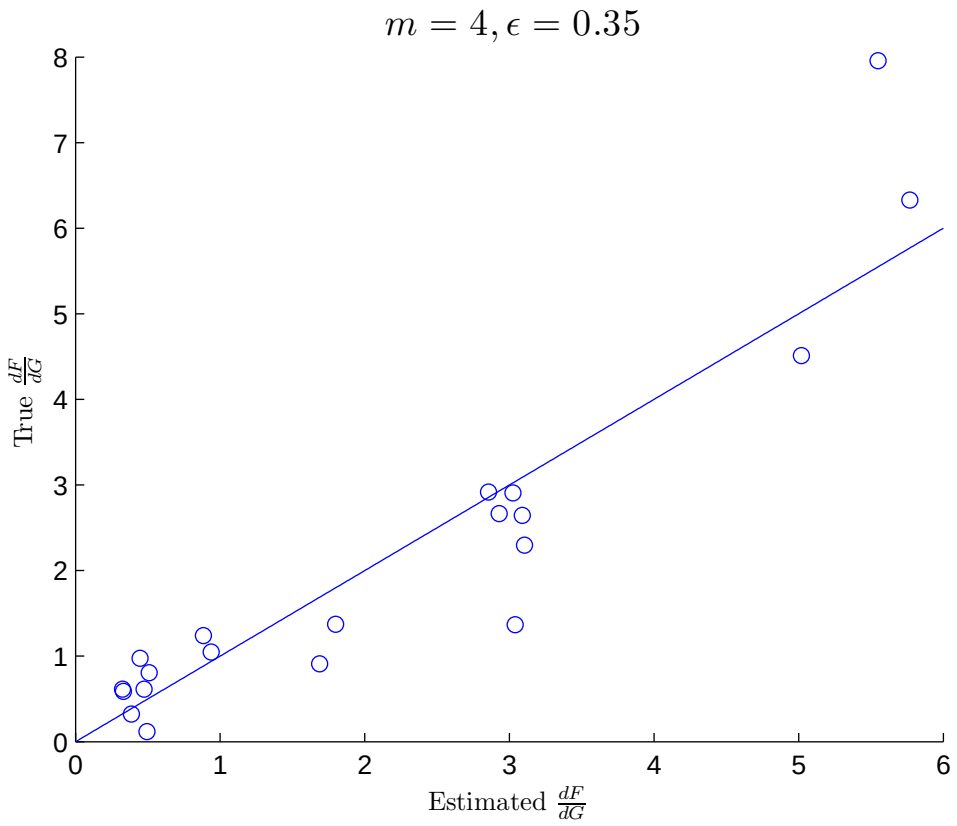


Figure 5.3: A plot of the true versus estimated values for dF/dG on the observed data Ω_S. The closer a point is to the line $y = x$, the better the density estimate at that point.

5.2 Tropical cyclone application

In this chapter, we apply the method of Chapter 4 to the tropical cyclone environment. To evaluate the performance of ABDE, we begin with a simulation study, sampling the observed data from a known distribution to which we can compare the estimated distribution produced by ABDE.

5.2.1 Simulation distribution

The distribution for both the data and the instrumental distribution will come from the same distribution family – an adaptation of a model from climate science, specifically that of Emanuel et al. (2006). It is one of the class described in Section 1.2 that treats a storm as a random walk over the ocean, coupled with a genesis distribution and a fairly complicated mechanism for lysis.

To formulate a data distribution for the simulation, the Emanuel et al. method is fit to the 409 observed tracks from HURDAT. This estimate produces realistic tracks and serves as F – the "truth" – for the simulation. From this F, a new set of 409 tracks are drawn, which we treat as the observed tracks for the purpose of this simulation study. To this new set, we re-fit the Emanuel et al. method to produce G, the instrumental distribution. It should be emphasized that there is no model uncertainty in the estimation of G — it is fit on the data from F assuming *the same underlying model that F comes from.* Figures 5.4 and 5.5 show four thousand tracks from this F and G, respectively.

This set-up mimics exactly the situation for which we feel ABDE would be useful: a scientist encounters a small set of high-dimensional data and fits a very large, interpretable

stochastic model to it. However, this model is not parsimonious enough to sufficiently accommodate a bias-variance tradeoff. Rather than do formal model selection on the input variables, our method uses that large model as is for the instrumental distribution and applies an ABDE layer on top to reduce the dimension of the problem.

Discussion of Emanuel et al.

A major difference between the distribution that we use and Emanuel et al. (2006) is that they are interested in a spatial-temporal estimate, whereas we are aggregating over time. In this discussion of Emanuel et al., we will elide the discussion of how they handle the time dimension because we are not replicating it. They begin by estimating the genesis distribution by taking a $0.5^\circ \times 0.5^\circ$ grid over the Atlantic and counting the number of occurrences within each grid. That histogram is then smoothed over using a Gaussian kernel which is isotropic in latitude and longitude but which varies as a function of location; specifically, the variance of the kernel is extended until a selected number of points are included or a 15° limit in latitude-longitude is exceeded. No genesis probability is assigned to land areas.

Once a lysis position is chosen for a track, it is developed as a sequence of 6-hour displacements. Emanuel et al. felt that it is more stable to model the differential movement of the tracks, so propagation is built up as a sequence of kernel density estimates, where the *change* in speed s and direction θ of the next step is conditional on the current speed and direction. In other words, the method first estimates the time homogenous series of conditional densities:

$$\mathbb{P}(\dot{s}_i, \dot{\theta}_i | s_{i-1}, \theta_{i-1}). \tag{5.1}$$

For a current speed and direction between the $(i-1)^{st}$ and i^{th} displacement location, s_{i-1} and θ_{i-1} respectively, a sample $\dot{s}_i, \dot{\theta}_i$ is drawn from the appropriate conditional density from Equation 5.1, and the $(i+1)^{st}$ displacement location is calculated:

$$\begin{pmatrix} x_{i+1} \\ \\ y_{i+1} \end{pmatrix} = \begin{pmatrix} x_i \\ \\ y_i \end{pmatrix} + (s_{i-1} + \dot{s}_i \delta t) \frac{180}{\pi a} \begin{bmatrix} \sin(\theta_{i-1} + \dot{\theta}_i \delta t) \\ \\ \frac{\cos(\theta_{i-1} + \dot{\theta}_i \delta t)}{\cos\left(\frac{\pi y_i}{180}\right)} \end{bmatrix}, \tag{5.2}$$

where a is the radius of the Earth and δt is six hours. Furthermore, Emanuel et al. assume conditional independence:

$$\mathbb{P}(\dot{s}_i, \dot{\theta}_i | s_{i-1}, \theta_{i-1}) = \mathbb{P}(\dot{s}_i | s_{i-1}) \cdot \mathbb{P}(\dot{\theta}_i | \theta_{i-1}). \tag{5.3}$$

The distribution we use in this thesis is essentially the same. The primary difference is in how to decide when the track should stop. Emanuel et al. use a two-pronged approach: the first is to construct an 8-step "sampling schedule" made up of various combinations of spatial and temporal resolutions for binning the data, and then stop the track when there is not sufficient data in any of the bins. However, they state that stopping for this criteria is very rare, and the primary reason that a track is stopped is because it enters an area of "weakly observed hurricane activity".

We also felt that tracks should stop when they enter a region of weakly observed hurricane activity, and so we construct a two-dimensional kernel density estimate over the

ocean, the observations for which are the endpoints of the regularized segments – not the tracks themselves. One can then use a threshold to construct a termination boundary and stop any track that ventures beyond the boundary. But tracks do not only stop because they have entered a dead zone— we see in the observed data many tracks which die right in the middle of a region of heavy TC activity. So we also fit a negative binomial distribution to the segments lengths of the observed data, and as part of the sampling process randomly select a segment length for each track. Then tracks die when either they enter the dead zone or they reach the determined length.

5.2.2 Density estimation

The distribution of the previous section is first estimated for the 409 observed tracks from HURDAT, a sample of which is shown in Figure 5.4. For the purposes of this simulation, that density estimate will serve as F, the underlying data distribution. From F, another 409 tracks are drawn as the observed data Ω_S. To construct an instrumental distribution G, we fit the method of Emanuel et al. to Ω_S; a sample from G is shown in Figure 5.5. We then apply ABDE to Ω_S and G to construct $\widehat{dF/dG}$. As in Chapter 3, the distance metric d for tracks is the sum of the distance between the regularized segment endpoints.

The question of interest is this: can ABDE do a better job of estimating expectations with respect to F than G can, even though G was fit with the same family as F? We assess the quality of the density estimation using both global and local checks – the former by plotting MISE as a function of dimension for our proposed method and the KDE method, and the latter by estimating a few particular integrals.

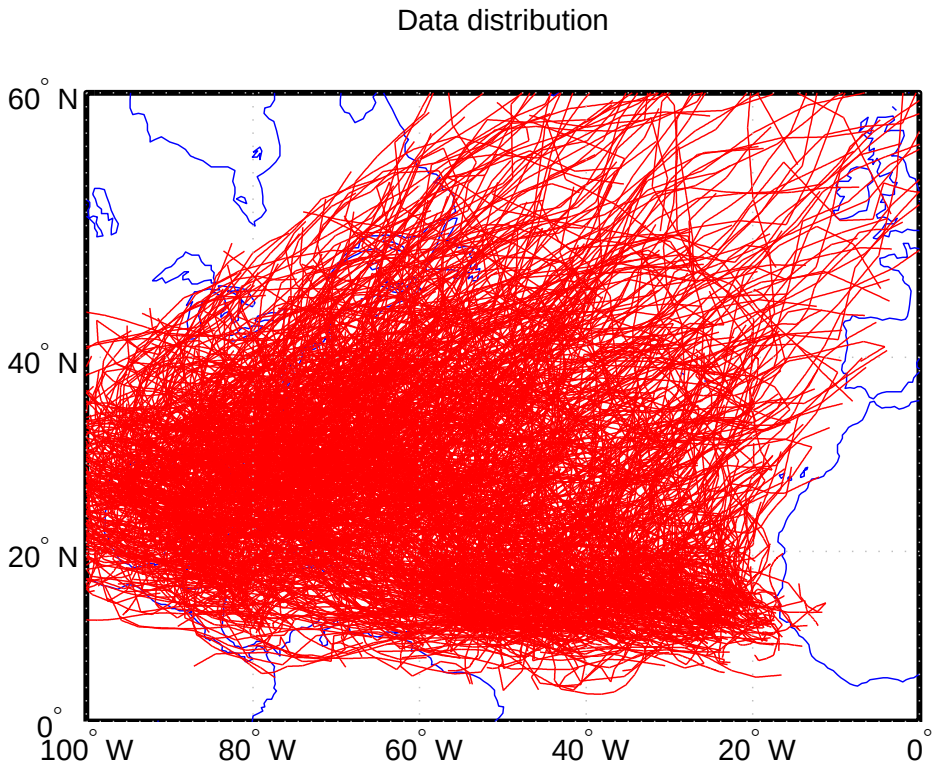


Figure 5.4: A sample of 4000 tracks drawn from data distribution F.

Figure 5.6 illustrates the MISE versus dimension for both the ABDE and the KDE methods. We see that for a data dimension of three, the MISE are quite similar, but as the dimension increases the ABDE does a better job of negotiating the bias-variance tradeoff.

The quality of Monte Carlo integrals for various particular "target functions" ($h(x)$ in Equation 4.1) can give a sense of performance on particular questions of scientific interest. The map shown in Figure 5.12 shows two regions — one in Africa, one in the Great Lakes region — which form the target functions for the importance sampling; $h(x)$ is simply an indicator function for whether track x ever pierces the region. On those two regions, we

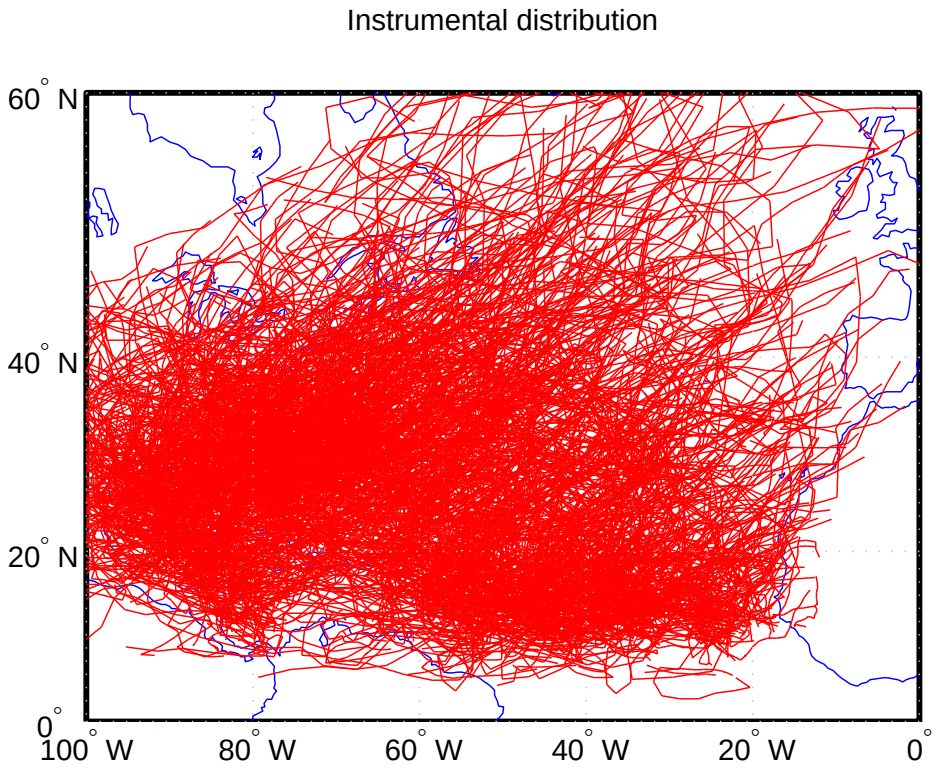


Figure 5.5: A sample of 4000 tracks drawn from instrumental distribution G.

compare:

- the true probability: $\mathbb{E}_F(h(X))$
- the probability under G using straight-forward Monte Carlo integration: $\mathbb{E}_F(h(X)) \approx \overline{h(X)}$ for $X \sim G$.
- the probability under G using the nonparametric importance sampling: $\mathbb{E}_F(h(X)) \approx \overline{\frac{h(X)g(X)}{j(X)}}$ for $X \sim J$ with density $dJ/d\mu = j(x)$, where J is chosen using the method of Zhang (1996).

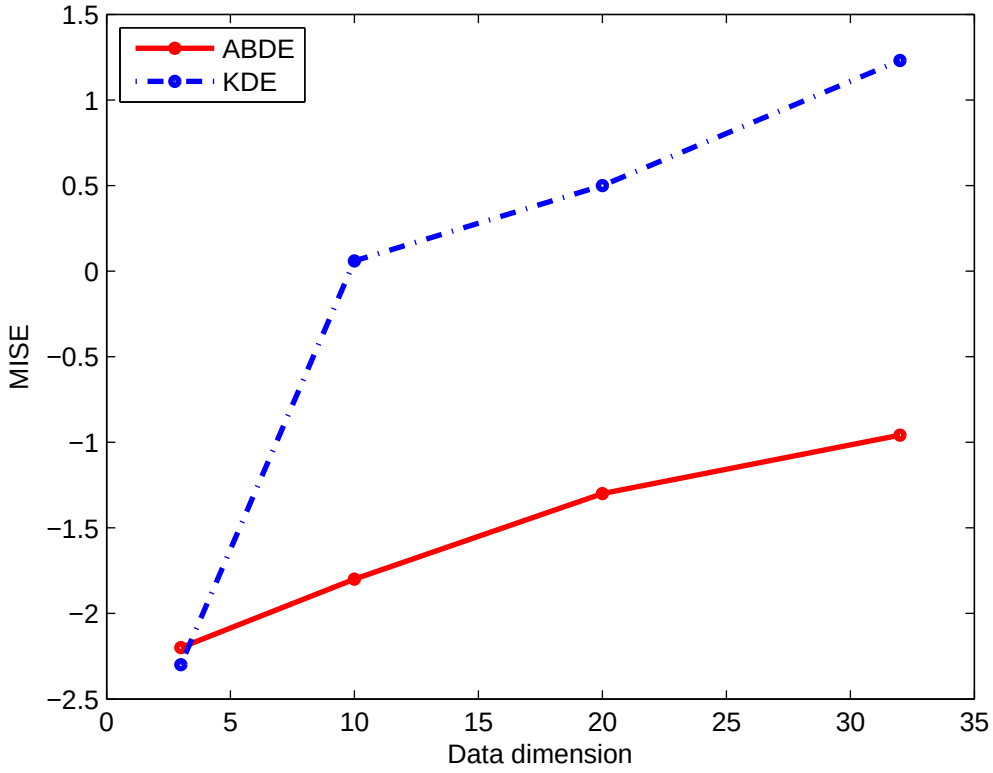


Figure 5.6: Mean integrated squared error versus dimension.

- the probability under ABDE: $\mathbb{E}_F(h(X)) \approx \overline{r(X)}$ for $X \sim G$.

To compare the estimation of the expected values, we repeated the simulation not once but twenty-five times. In each case the F estimated from the HURDAT data remains the same, but a new Ω_i for $i \in \{1, \ldots, 25\}$ of size 409 was drawn and the ABDE applied to each Ω_i separately. In 21 of the simulations, $q = 3$ is the optimal number of basis functions, whereas in four simulations $q = 4$ is chosen. For both regions, we can see in Figure 5.13 that G under both Monte Carlo integration and importance sampling over-estimate the probability and the ABDE adjusted the G to move the estimate substantially closer to the

true probability.

A sample of tracks from the instrumental distribution, colored by their estimated density $\widehat{dF/dG}$, is show in Figure 5.7 through Figure 5.11. It demonstrates what can be seen qualitatively from Figures 5.4 and 5.5, namely that the instrumental distribution is more diffuse than the true data distribution, and as a result higher density is assigned to more central tracks.

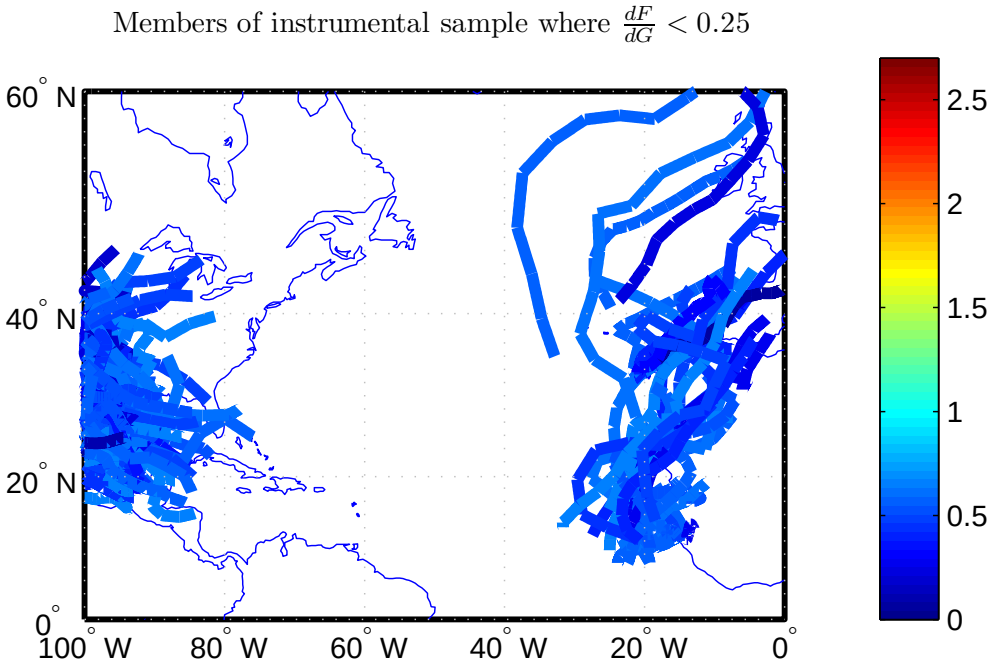


Figure 5.7: Members of instrumental sample where $\frac{dF}{dG} < 0.25$

These promising initial results would likely improve if the criterion being optimized

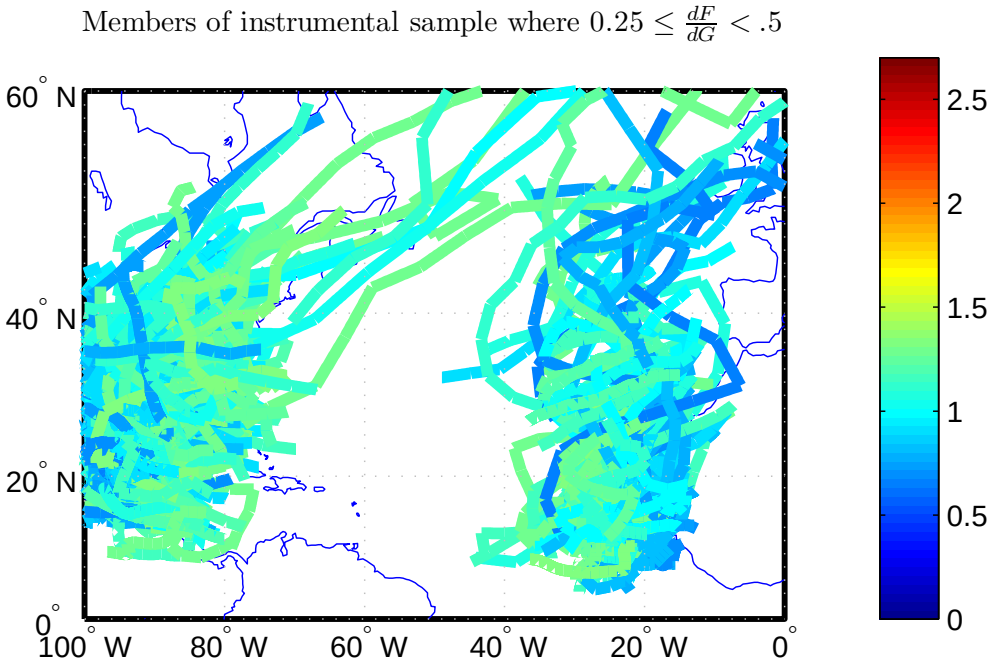


Figure 5.8: Members of instrumental sample where $0.25 \leq \frac{dF}{dG} < .5$

over were tailored towards the estimation of such probabilities. Thus far, the loss function under consideration, the MISE, penalizes the estimate very little for (relatively) large errors in regions of small probability under F; this is simply a property of integrating squared errors in the density. Other choices for the loss function could be justified, including ones which penalize the estimator based on the size of the error relative to the true density in the region.

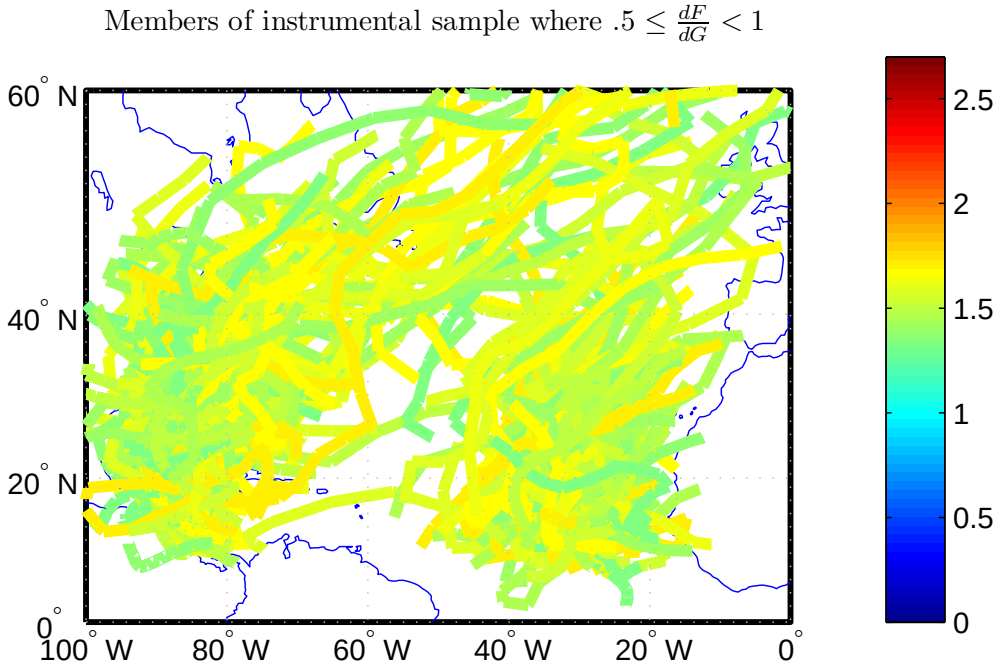


Figure 5.9: Members of instrumental sample where $0.5 \leq \frac{dF}{dG} < 1$

5.3 Density estimation for TCs

In the previous section, we worked with the TC data as the basis for a simulation example; now we will investigate the unknown data density of the TCs. In other words, Ω will now be the observed TCs from HURDAT, from unknown data distribution F, and what served as F in the previous section — a fit of the method of Emanuel et al. to Ω — will serve here as G, the instrumental distribution. The ABDE fit selects $q = 3$ and $\epsilon = 183,200$. Unfortunately visualization of the full high-dimensional density is not possible, so we sampled 100,000

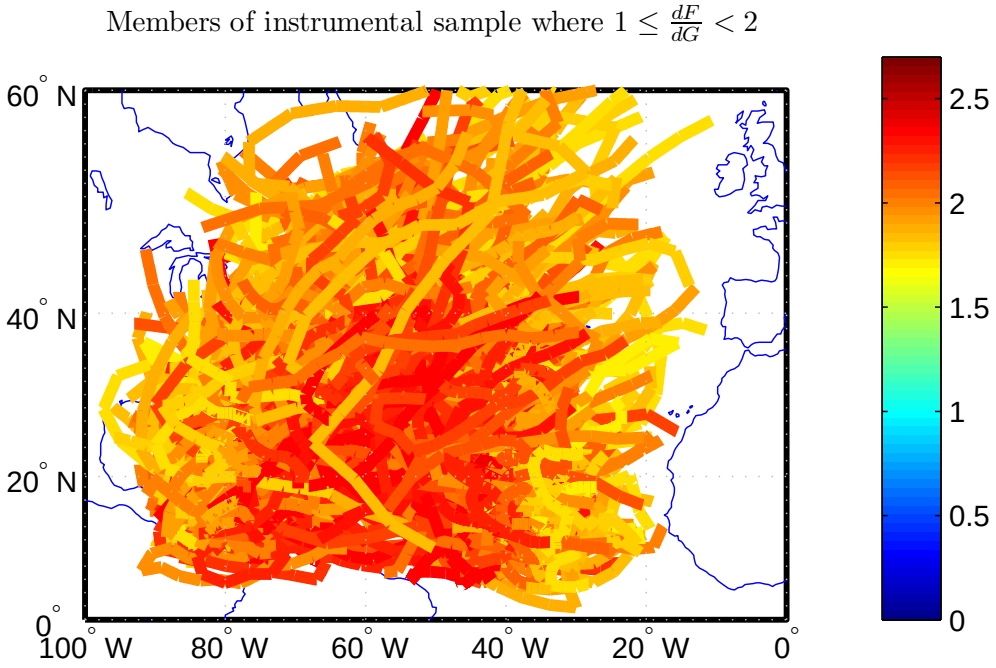


Figure 5.10: Members of instrumental sample where $1 \leq \frac{dF}{dG} < 2$

tracks from G and the 10 with the highest and lowest estimated densities dF/dG are shown in Figures 5.14, 5.15, and 5.16.

5.4 Summary

In this chapter, we saw two applications of ABDE. In both cases, ABDE in combination with importance sampling does a better job of estimating integrals with respect to the true data distribution F than kernel density estimation alone.

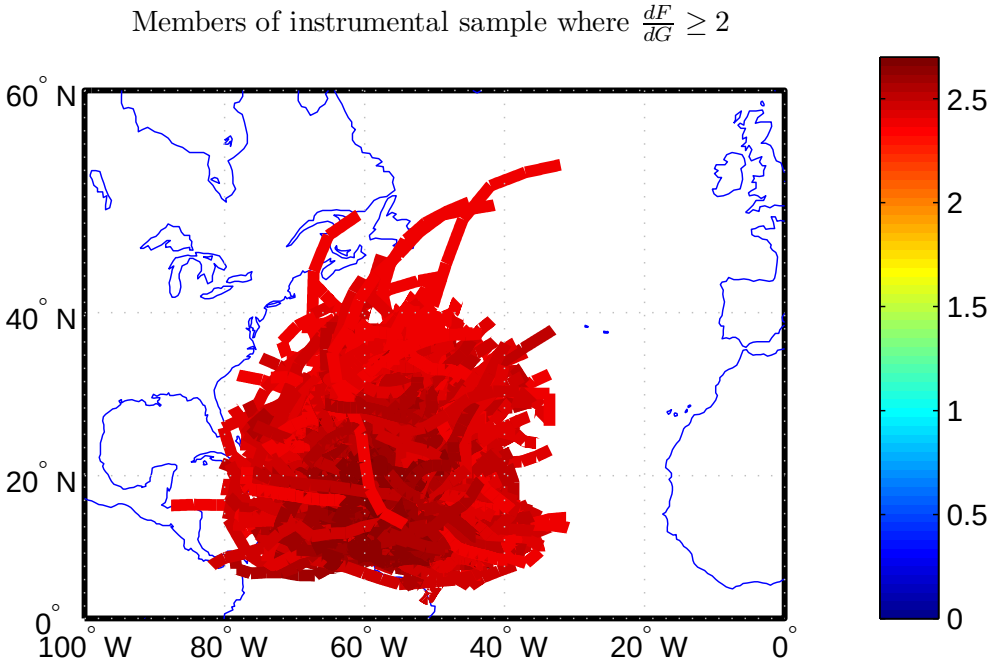


Figure 5.11: Members of instrumental sample where $\frac{dF}{dG} \geq 2$

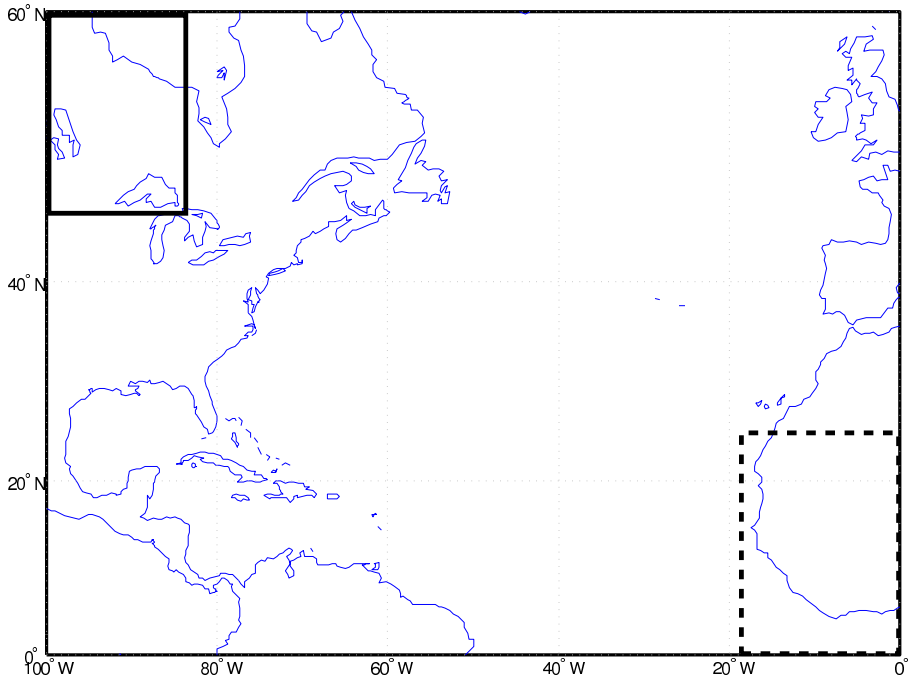


Figure 5.12: The solid rectangle in the upper-left corner of the graph delineates the Great Lakes region of interest; the dashed rectangle delineates the African region.

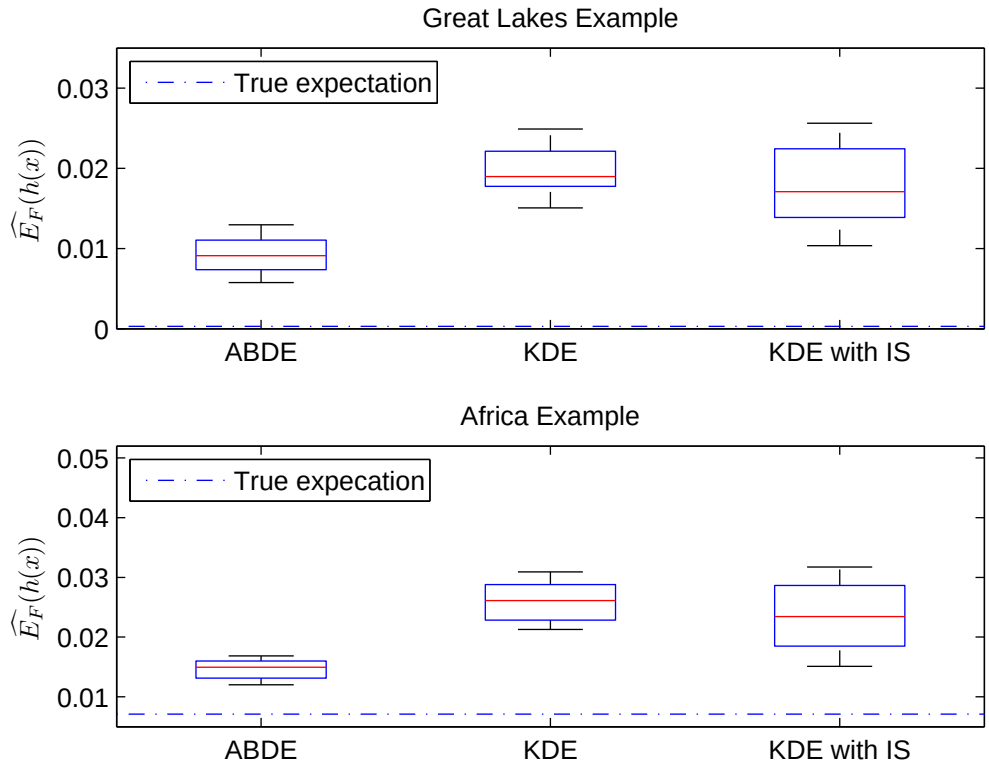


Figure 5.13: A comparison of performance by region. Each box plot corresponds to 25 simulations – each of the 25 begins with 409 different "observed" tracks, although all are drawn from the same distribution.

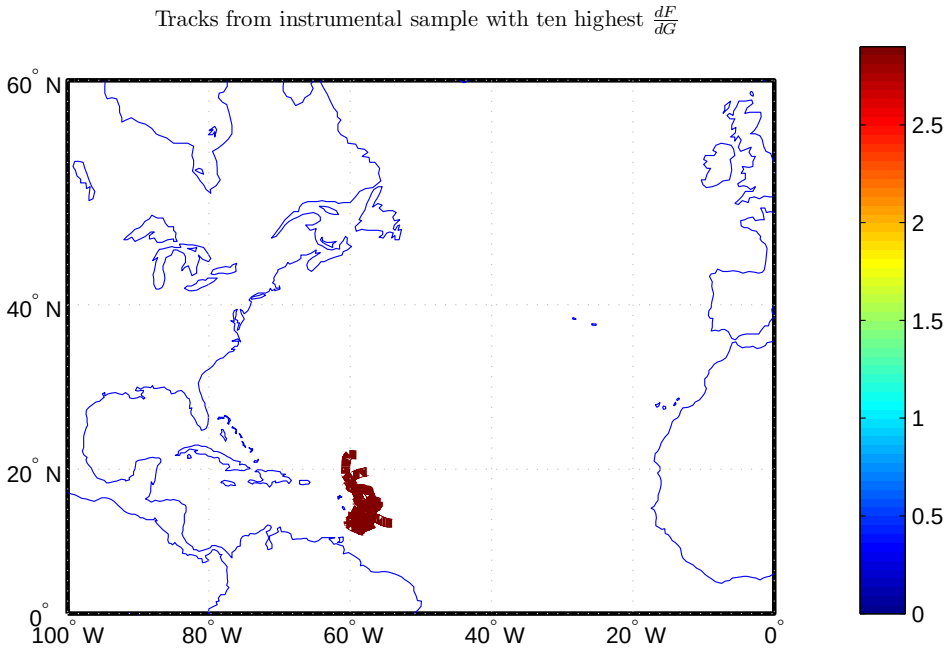


Figure 5.14: Of a large instrumental sample, the ten tracks with the highest values of dF/dG, corresponding to those tracks whose probability is highly underestimated by kernel density estimation.

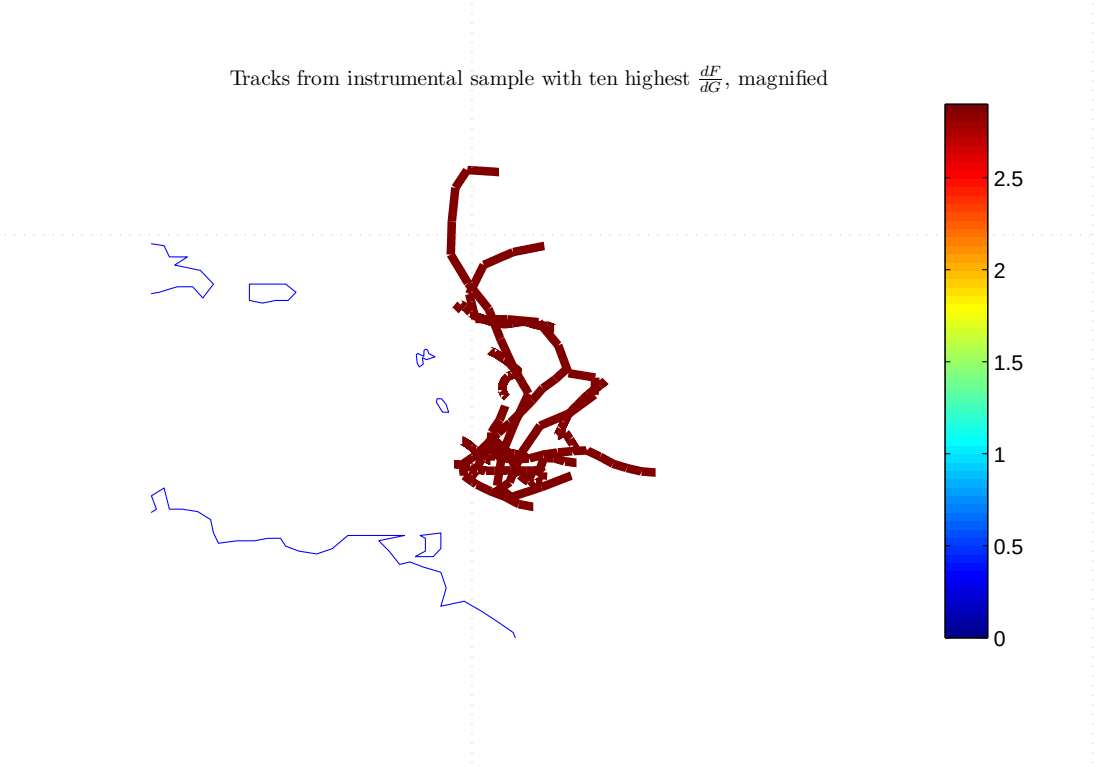


Figure 5.15: A magnified view of Figure 5.14

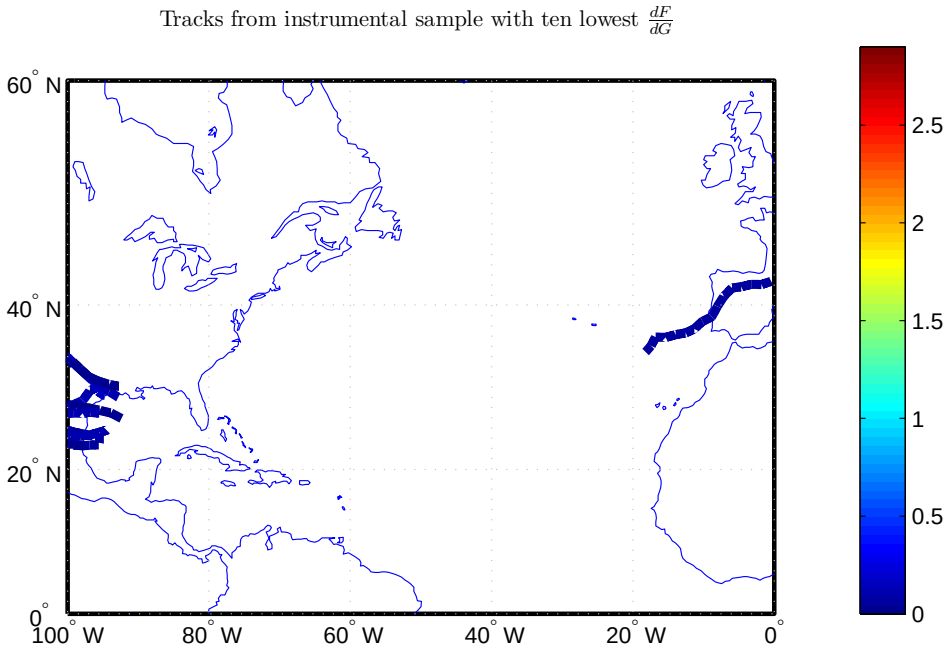


Figure 5.16: Of a large instrumental sample, the ten tracks with the lowest values of dF/dG, corresponding to those tracks whose probability is highly overestimated by kernel density estimation.

Chapter 6

Potential continuations and conclusions

One question of strong interest is whether the spatial variation of TCs is a function of other climate variables, such as sea surface temperature, the El Niño-Southern Oscillation, and the North Atlantic Oscillation (Elsner and Kocher, 2000; Saunders and Lea, 2008). One natural way to approach this is as a conditional density estimation problem, in which the distribution over track space changes with the values of the climate variables. In other words, consider not just $f(X)$ but $f(X|SST = y)$. However, this leaves out considerations of frequency, which are definitely of interest among the climate science community (Saunders and Lea, 2008; Smith et al., 2010; Webster et al., 2005; Goldenberg et al., 2001). For example, it might be the case that a one-degree increase in SST does not change just the *distribution* of tracks, but also change the expected *frequency* — certainly both facets are of interest here. This leads us to pursue models beyond density estimates, such as

spatial point processes with covariate data.

In addition to extending the spatial point process work, there are four main avenues under consideration for extending the work in this dissertation. Namely, developing a more formal understanding of the assumptions under which this method will work and establishing consistency results; studying how the smoothing effect of increasing ϵ and the smoothing effect of decreasing q affect each other; adjusting for the presence of ϵ in plug-in methods for selecting q; and improving our validation methods for high-dimensional data.

6.1 Climatic predictors

Some of the most important questions regarding TCs could be addressed, at least partially, through a better understanding of the relationship between TC occurrence and other measurable characteristics of the climate system. Such relationships could be utilized in, for instance, creating and verifying complex simulation models, predicting future trends in TC activity, and understanding human influence on the climate system. Specifically, an area of great concern is the effect that rising sea surface temperatures (SST) might have on the frequency and/or intensity of TCs.

First consider a set-up similar to Hall and Jewson (2008): they focused on the 19 hottest and 19 coldest years from 1950 to 2005, where a year's "temperature" was defined as the July-August-September SST averaged over a region of the Atlantic. After dividing the North American continental coastline into six major segments — the U.S. Northeast, the U.S. Mid-Atlantic, Florida, the U.S. Gulf, the Mexican Gulf, and the Yucatan Peninsula — they performed hypothesis tests on the difference in yearly landfall rates between the

hot and cold years. In all regions but the U.S. Northeast, the landfall rate was higher in hot years, with the difference in the Yucatan being found as statistically significant.

Their approach requires that one assume a particular theory about the relationship between SST and landfall rates. It also requires that the coastline be divided into somewhat arbitrary, large segments. It would be preferable to have the densities inform us as to the regions which are experiencing differences among the hot and cold years. For example, consider Figure 6.1, which shows the density estimates found when our methods are applied separately to the tracks from cold years (Figure 6.1(a)) and the tracks from hot years (Figure 6.1(b)). If we focus on the regions in which the hot density is much higher than the cold density — specifically, the range $D1 \in [2.4, 3], D2 \in [-.9, -.4]$, found using the technique of Section 3.2 — we can map the tracks that fall into that region, as shown in Figures 6.2(a) and 6.2(b). We see that most of these tracks are southern U.S. and Central American landfalling tracks, which comports with Hall and Jewson (2008). Density estimation allows us not only to test hypotheses about the effect of climatic predictors on TCs, but also provides a way of generating insight into the nature of these relationships.

While the above method allows for more flexibility than cutting a coastline up a priori, it is forced to reduce SST to a binary variable: "hot year" or "cold year." These results treat a 56-year stretch of extreme climate events as samples from two distributions. It is desirable to extend this type of approach to a more granular specification of the climatic predictors.

6.1.1 Discussion of spatial point process models

Spatial point processes are point processes in two or more dimensions (Baddeley, 2008; Møller and Waagepetersen, 2006). They are treated as a special case because the lack of ordering in multi-dimensional spaces means that many metrics for one dimensional point processes do not apply, i.e. inter-arrival times, as in when the one dimension is time.

A spatial point pattern $\{x_1, x_2, \ldots, x_m\}$ of points contained in $S \subset \mathbb{R}^\ell$ is a realization of a finite random subset of S, called a spatial point process X. As in the previous chapters, we will model the point process over diffusion space, not track space, so ℓ will not be high dimensional. $N(B) = n(X_B)$, where $X_B = X \cap B$, is the associated random variable containing the number of points. As in point processes, the goal with a spatial point process is to estimate the moment measure ν, the mean number of points for any bounded B:

$$\nu(B) = EN(B), \tag{6.1}$$

which, analogously to density estimation, is modelled as a function of the *intensity function* λ:

$$\nu(B) = \int_B \lambda(u)du \tag{6.2}$$

where $\lambda(u) \geq 0 \; \forall \; u$.

Thus, rather than estimating f, the density function, we are estimating λ, the intensity function, which need not integrate to one.

A spatial Poisson process assumes that the members of the point pattern occur inde-

pendently of one another; however an intensity function can be quite complicated and can account for interaction between the points. For example, spatial inhibition is a sort of repulsion effect for points of the pattern (for example, if modeling the location of trees, because of resource competition and lack of sunlight, two trees will not have trunk centers that are less than a few inches from each other). Spatial aggregation is the other extreme, and applies to cases when points are more likely to cluster together than under independence. In the context of real-time or short-term hurricane forecasting, one would need to model something like aggregation (two storms cannot occur too closely at the same time before they become one storm). However, because we are taking a long-term ensemble approach, the assumption of independence is reasonable.

Spatial point patterns can also contain *marks*, which are additional response information (e.g. in addition to the location of a storm, we might also record the amount of property damage it caused) and *covariates*, which is associated data that is treated as an explanatory variable (e.g. SST). If the covariate data is potentially observable at all spatial locations $x \in \mathcal{X}$, it is a *spatial function* $Z(u)$. The goal of this chapter, of course, is to treat the covariate data as spatial covariate functions.

But we must decide how to assign a single value of a climate variable to an entire high-dimensional track — over the course of its life, a TC will be at many positions at many times, each with a different SST. Do we select SST at the genesis point? Maximum SST over the track? Some sort of distance- or time-weighted average? We discuss the approach that we use for this problem in Section 6.1.3, but the answer will vary depending on context.

6.1.2 Point process model

We select an spatial inhomogeneous Poisson process that is a function of both the path the track takes and an averaged SST:

$$\log \lambda(u,t) = \alpha_0 + \alpha_1 \psi_1(u) + \alpha_2 \psi_2(u) + \alpha_3 s(u,t) \tag{6.3}$$

where u is the track, t is the month, and $s(u,t)$ is a function for the integrated SST of a track with spatial configuration u in month t. The model is fit using the R package `spatstat` (Baddeley and Turner, 2005). This method requires that the covariates can be measured at places of non-occurrence, not just at the observed data points. This is not a problem for some covariates of interest for this application – integrated SST, or presence of El Niño – but is for others – for example, the intensity (max wind speed) of a storm that did not occur could only be estimated with a very detailed climate model (and while there are plenty to choose from, they often present conflicting intensity estimates). The selection of $d = 2$ was made ad hoc.

6.1.3 Covariate function

The form of $s(u,t)$ used in this initial project averages SST over the track. The SSTs are themselves estimates – monthly average temperatures on a $5^\circ \times 5^\circ$ grid over the ocean. Each track u is regularized to n equally-spaced points, and the SST of the closest grid point during the month t that the TC occurred is assigned to each track point.

6.1.4 Results

The fitted model is

$$\log \lambda(u,t) = 3.81 + 0.114\psi_1(u) - 0.063\psi_2(u) + .462s(u,t). \tag{6.4}$$

One implication that follows from the estimate is that, should the SST for the entire ocean rise $1°C$ next year, we would expect exp(.462) times as many storms – roughly 59% more. This number might seem large, but is not out of line with other climate results. Saunders and Lea (2008) found that "increasing the August-September SST in the [main development region, an area where many TCs begin] by $0.5°C$ above its climate norm value of $27.3°C$ is linked to increases above the 1950-200 norm values of $31\pm$ 17 %" in the number of TCs.

However, the model of Equation 6.3 is more detailed than just integrating over the whole space to produce count data. The multiplicative effect on $\lambda(u,t)$ of a one-degree change in $s(u,t)$ means that the absolute change will be more pronounced for tracks whose averaged SST is already on the high side. The model could be extended to allow for interactions between the ψs and $s(u,t)$ as well. Figure 6.3 shows a sample of tracks colored by their estimated intensity.

6.1.5 Extensions to the spatial point process

The two most important extensions to the spatial point process work would methodically choose the number of diffusion coordinates, and adding more climatic predictors. The

work on goodness-of-fit tests for spatial point processes is not as developed as for density estimation, but there has been recent interest. Baddeley et al. (2005) define and investigate residuals for a point process model, along with diagnostic plots. As useful as these are, they do not provide as formal a comparison between models as something like a mean integrated square error. Guan (2008) proposes a new goodness-of-fit test for inhomogeneous spatial Poisson process.

6.2 Underlying assumptions and asymptotics

There are several outstanding questions about the underlying assumptions and asymptotics of the methods proposed in this dissertation. What, exactly, does it mean for an instrumental distribution G to be "good", and what is the relationship between a measure of its goodness and asymptotic properties of the estimator? What is the effect of having not a true orthonormal basis but an approximate one? And why, from a qualitative perspective, does the method work in high dimensions?

6.2.1 Sufficient properties of G

Orthogonal series density estimators for a density $g(x)$ on n observations, truncated to $q(n)$ terms with estimates for α chosen as in Equation 4.10, can be shown to converge in MISE as $n \to \infty$ assuming that $g(x)$ is square integrable and $q(n)/n \to 0$ as $n \to \infty$ (Schwartz, 1967). Under various sets of tighter assumptions, one can establish different rates of MISE convergence or stronger forms of convergence (Schwartz, 1967; Ahmad, 1982; Hall, 1986).

We suspect the careful selection of G can improve those asymptotic results, but our

method is, for now, heuristically motivated. We only require that G be a dominating measure for F and that G be "close enough" to F.

For example, in Equation 4.12, we provided a (loose) bound on $\widehat{f}$. The existence of a bound is useful for rejection sampling, but also limits the scope of the method; if the true density exceeds the bound in some region, we will not estimate it well there.

Obviously, the smoother F is with respect to G, the better, as the density will require fewer terms to produce the same bias; for example, we may choose to characterize the rate of convergence by the supremum of the second derivative of dF/dG, or by the second derivative's L_2-norm.

6.2.2 Approximately orthonormal basis

Even having chosen a sufficient G, the basis used is only approximately orthonormal with respect to S_ϵ; it is crafted from a very large, but finite, sample from G, and the approximate eigenfunctions are found using the Nyström extension.

The error that this introduces, and its effect on convergence rates, will be a function of G and of the instrumental sample size used to construct the approximate and requires further investigation.

Qualitative understanding

Another aim is to understand more intuitively why the method works. An obvious component is that by assuming the existence of a sufficient instrumental distribution, we are assuming access to additional knowledge about the data distribution; any method that

makes reasonable use of that knowledge will perform better than if it did not. We have chosen this method in large part *because* it provides such a natural way to take advantage of G.

Furthermore, the adaptive basis generated by the diffusion map allows for the generation of a basis on a high-dimensional space without the need for a basis for each dimension, interaction terms, etc.

6.3 Smoothing and ϵ versus q

As discussed above, increasing ϵ has a smoothing effect, whereas decreasing q leads to smoothing. Suppose that for a particular ϵ and q, we determine that the model is over-fitting. Is there any way to understand a priori in what cases we might want to correct the over-smoothing by adjusting ϵ versus adjusting q? Are adjustments to the two, but for the discrete nature of q, interchangeable?

For example, if for a particular ϵ_0 it is determined that the plug-in optimal truncation point is $q^*_{\epsilon_0}$ — meaning the next term increases the variance by more than it decreases the squared bias — can we establish that, for $\epsilon_1 < \epsilon_0$, $q^*_{\epsilon_1} \geq q^*_{\epsilon_0}$?

6.4 Adapt plug-in methods to account for ϵ

As discussed in Section 4.6, the ϵ-dependent basis makes this work different from traditional orthogonal series density estimation. Assuming a fixed basis $\{\phi_i, i \geq 0\}$, the conventional approach to selecting q was popularized by Kronmal and Tarter (1968) and entails stopping

at the first basis function j for which

$$\frac{2}{(n+1)}\frac{1}{n}\sum_{i=1}^{n}\phi_j^2(X_i) > \left(\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i)\right)^2. \tag{6.5}$$

There have of course been extensions to this method. Diggle and Hall (1986) present a method that does not consider each term individually, but considers the entire sum up to j in determining whether to stop at j. Hart (1985) and Efromovich (1999) also present extensions.

One could consider whether the method of analytically selecting a truncation point can be adapted to include ϵ and avoid a CV search over the two parameters. A straightforward way to do this would be to only search over ϵ and, for a given ϵ, use the plug-in method to select q. However, the output of the proposed work of Section 6.3 could aid in producing even more sophisticated plug-in methods.

6.5 Validation methods

The validation technique in Section 3.2 is currently heuristically motivated. It would be desirable to establish consistency results for the method; for example, to show that the adaptive basis density estimate $\widehat{f}$, constructed with a particular distance metric d, is consistent with respect to d.

6.6 Conclusions

The main goal of this thesis was to explore the potential of diffusion maps in nonparametric density estimation. The method's success in other high-dimensional problems such as regression and clustering made it a prime candidate for density estimation. We pursued the idea of using importance sampling to create high-dimensional densities not with respect to the Lebesgue measure, but with respect to other high dimensional probability densities.

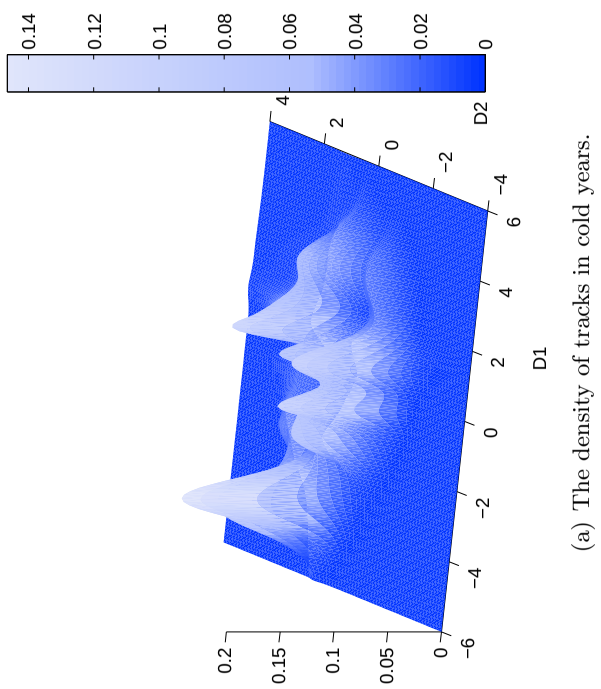
The main findings of this dissertation are:

1. that the bases produced by diffusion maps are very well-suited for use in orthogonal series density estimation;

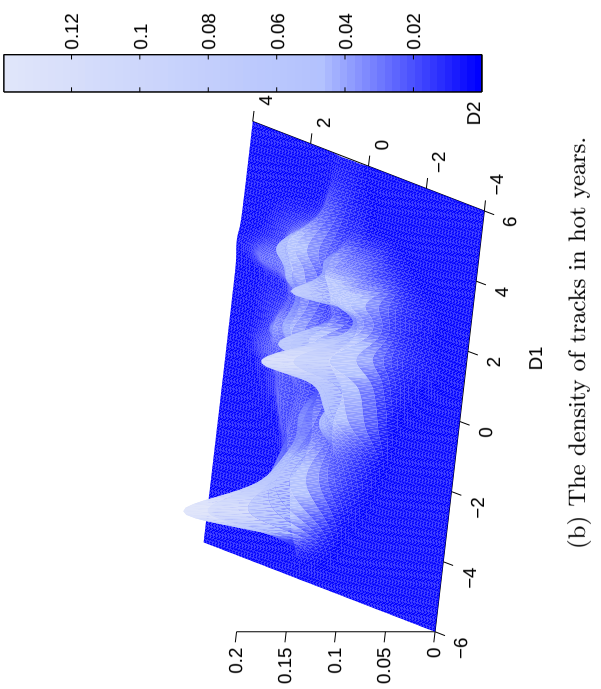
2. that treating an existing distribution for a particular domain — which may integrate crucial scientific/domain knowledge but not account for high dimensionality — as an instrumental distribution in importance sampling is a straightforward way to integrate that information without relying on overly ad hoc adjustments;

3. an application of this method to the tropical cyclone context;

4. and that the method can be extended to use climatic variables as spatial covariates.

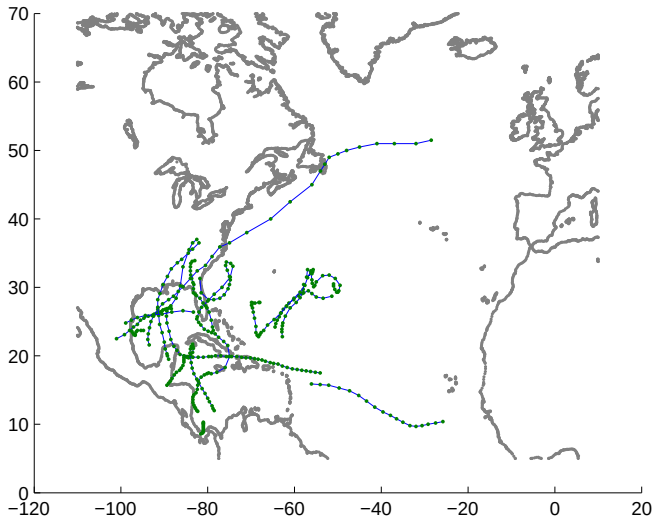


(a) The density of tracks in cold years.

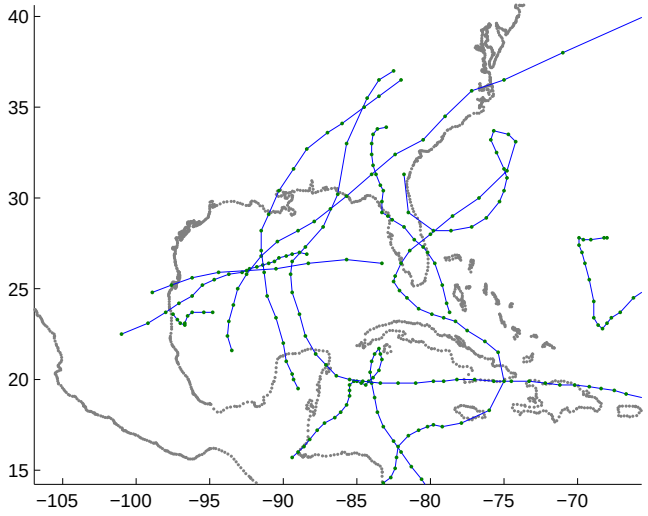


(b) The density of tracks in hot years.

Figure 6.1: Densities for tracks conditioned on hot and cold years.



(a) The tracks in the discrepancy region.



(b) A closer look at the tracks.

Figure 6.2: Tracks from the region where the density is much higher in the hot years.

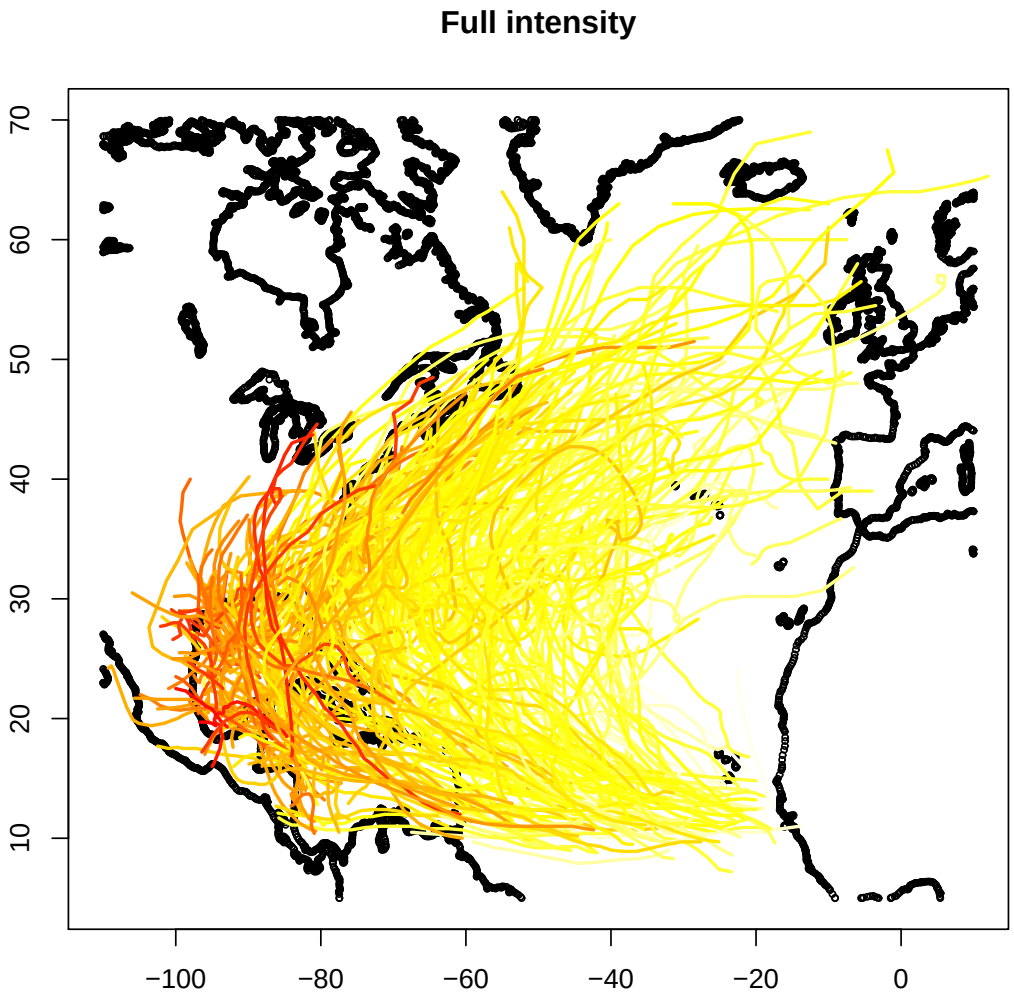


Figure 6.3: Tracks colored by their intensity as estimated by Equation 6.4.

Bibliography

Ahmad, I. A. (1982). Integrated mean square properties of density estimation by orthogonal series methods for dependent variables. *Annals of the Institute of Statistical Mathematics*, 34:339–350.

Annan, J. and Hargreaves, J. (2007). Efficient estimation and ensemble generation in climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):2077–2088.

Baddeley, A. (2008). Analysing spatial point patterns in r. *Workshop on Analysing spatial point patterns.* Version 3, `http://www.stats.uwo.ca/faculty/kulperger/S9934a/Computing/Spatstat-pn0y.pdf`.

Baddeley, A. and Turner, R. (2005). Spatstat: An r package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42.

Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Belkin, M. and Niyogi, P. (2004). Semi-supervised learning on riemannian manifolds. In *Machine Learning*, pages 209–239.

Bell, G. D. and Chelliah, M. (2006). Leading tropical modes associated with interannual and multidecadal fluctuations in north atlantic hurricane activity. *Journal of Climate*, 19:590–612.

Bengtsson, T., Bickel, P., and Li, B. (2008). Curse-of-dimensionality revisited: Collapse of importance sampling in very large scale systems. In Nolan, D. and Speed, T., editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 2, pages 316–334. Institute of Mathematical Statistics.

Bengtsson, T., Snyder, C., and Nychka, D. (2003). Toward a nonlinear ensemble filter for high dimensional systems. *Journal of Geophysical Research*, 108.

Berliner, L. M. (2001). Monte carlo based ensemble forecasting. *Statistics and Computing*, 11(3):269–275.

Buchman, S. M., Lee, A. B., and Schafer, C. M. (2011). High-dimensional density estimation via sca: An example in the modelling of hurricane tracks. *Statistical Methodology*, 8(1):18–30.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Sparse density estimation with

l1 penalties. In *Proceedings of the 20th annual conference on Learning theory*, COLT'07, pages 530–543, Berlin, Heidelberg. Springer-Verlag.

Campbell, K. (2006). Statistical calibration of computer simulations. *Reliability Engineering & System Safety*, 91(10-11):1358 – 1363. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004) - SAMO 2004.

Coifman, R., Lafon, S., Lee, A., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. (2005). Geometric diffusions as a tool for harmonics analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431.

Crain, B. R. (1973). A note on density estimation using orthogonal expansions. *Journal of the American Statistical Association*, 68(344):964–965.

Diggle, P. J. and Hall, P. (1986). The selection of terms in an orthogonal series density estimator. *Journal of the American Statistical Association*, 81(393):230–233.

Donoho, D. L. (2000). Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality.

Efromovich, S. (1999). *Nonparametric Curve Estimation*. Springer, New York.

Elsner, J. B. (2006). Evidence in Support of the Climate Change-Atlantic Hurricane Hypothesis. *Geophysical Research Letters*, 33(L16705).

Elsner, J. B. and Kocher, B. (2000). Global tropical cyclone activity: A link to the north atlantic oscillation. *Geophysical Research Letters*, 27(1):129–132.

Emanuel, K., Ravela, S., Vivant, E., and Risi, C. (2006). A statistical deterministic approach to hurricane risk assessment. *Bulletin of the American Meteorological Society*, pages 299–312.

Freeman, P. E., Newman, J. A., Lee, A. B., Richards, J. W., and Schafer, C. M. (2009). Photometric redshift estimation using spectral connectivity analysis. *Monthly Notices of the Royal Astronomical Society*, 398:2012–2021.

Girolami, M. (2002). Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14(3):669–688.

Goldenberg, S. B., Landsea, C. W., Mestas-Nunez, A. M., and Gray, W. M. (2001). The Recent Increase in Atlantic Hurricane Activity: Causes and Implications. *Science*, 293(5529):474–479.

Guan, Y. (26 December 2008). A goodness-of-fit test for inhomogeneous spatial poisson processes. *Biometrika*, 95:831–845(15).

Hall, P. (1986). On the rate of convergence of orthogonal series density estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(1):115–122.

Hall, P. and Heckman, N. E. (2002). Estimating and depicting the structure of a distribution of random functions. *Biometrika*, 89(1):145–158.

Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374.

Hall, T. and Jewson, S. (2008). SST and North American Tropical Cyclone Landfall: A Statistical Modeling Study. `arXiv:0801.1013v1 [physics.ao-ph]`.

Hall, T. M. and Jewson, S. (2007). Statistical modelling of North Atlantic tropical cyclone tracks. *Tellus*, pages 486–498.

Hart, J. D. (1985). On the choice of a truncation point in fourier series density estimation. *Journal of Statistical Computation and Simulation*, 21(2):95–116.

Holland, G. J. and Webster, P. J. (2007). Heightened tropical cyclone activity in the North Atlantic: natural variability or climate trend? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1860):2695–2716.

Jarvinen, B. R., Neumann, C. J., and Davis, M. A. S. (1984). A tropical cyclone data tape for the North Atlantic Basine, 1886-1983, contents, limitations, and uses. *Technical Report NWS NHC 22, NOAA Technical Memo.*

Justel, A., Peña, D., and Zamar, R. (1997). A multivariate kolmogorov-smirnov test of goodness of fit. *Statistics and Probability Letters*, 35:251–259.

Kaplan, A., Cane, M. A., Kushnir, Y., Clement, A. C., Blumenthal, M. B., and Rajagopalan, B. (1998). Analyses of global sea surface temperature 1856-1991. *Journal of Geophysical Research*, 103:18567–18590.

Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by fourier series methods. *Journal of the American Statistical Association*, 63(323):925–952.

Kwok, J. T. Y. and Tsang, I. W. H. (2004). The pre-image problem in kernel methods. *Neural Networks, IEEE Transactions on*, 15(6):1517–1525.

Lafon, S. and Lee, A. B. (2006). Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(9):1393–1403.

Lee, A. B. and Wasserman, L. (2010). Spectral Connectivity Analysis. *Journal of the American Statistical Association*, 105(491):1241–1255.

Levina, E. and Bickel, P. J. (2004). Maximum likelihood estimation of intrinsic dimension. In *NIPS*.

Liu, H., Lafferty, J., and Wasserman, L. (2007). Sparse nonparametric density estimation in high dimensions using the rodeo. In Meila, M. and Shen, X., editors, *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, corrected edition.

Mann, M. E. and Emanuel, K. A. (2006). Atlantic hurrican trends linked to climate change. *EOS*, 87(24):233–244.

Mika, S., Schölkopf, B., Smola, A. J., Müller, K.-R., Scholz, M., and Rätsch, G. (1999). Kernel PCA and de–noising in feature spaces. In Kearns, M. S., Solla, S. A., and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11*. MIT Press.

Møller, J. and Waagepetersen, R. P. (2006). Modern statistics for spatial point processes.

National Hurricane Center (2009). Technical summary of the national hurricane center track and intensity models. `http://www.nhc.noaa.gov/modelsummary.shtml`.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press.

Oh, M.-S. and Berger, J. O. (1993). Integration of multimodal functions by monte carlo importance sampling. *Journal of the American Statistical Association*, 88(422):450–456.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering & System Safety*, 91(10-11):1290 – 1300. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004) - SAMO 2004.

Oouchi, K., Yoshimura, J., Yoshimura, H., Mizuta, R., Kusunoki, S., and Noda, A. (2006). Tropical cyclone climatology in a global-warming climate as simulated in a 20 km-mesh global atmospheric model: Frequency and wind intensity analyses. *Journal of the Meterological Society of Japan*, 84(2):259–276.

Richards, J. W., Freeman, P. E., Lee, A. B., and Schafer, C. M. (2009). Exploiting Low-Dimensional Structure in Astronomical Spectra. *Astrophysical Journal*, 691:32–42.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, 2 edition.

Rumpf, J., Weindl, H., Höppe, P., Rauch, E., and Schmidt, V. (2007). Statistical modelling of tropical cyclone tracks. *Mathematical Methods of Operations Research*, (3):475–490.

Saunders, M. A. and Lea, A. S. (2005). Seasonal prediction of hurricane activity reaching the coast of the united states. *Nature*, 434(21):1005–1008.

Saunders, M. A. and Lea, A. S. (2008). Large contribution of sea surface warming to recent increase in atlantic hurricane activity. *Nature*, 451(31):557–560.

Schwartz, S. C. (1967). Estimation of probability density by an orthogonal series. *The Annals of Mathematical Statistics*, 38:1261–1265.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.

Sinnott, R. W. (1984). Virtues of the haversine. *Sky and Telescope*, (2).

Smith, D. M., Eade, R., Dunstone, N. J., Federday, D., Murphy, J. M., Pohlmann, H., and Scaife, A. A. (2010). Skilful multi-year predictions of atlantic hurricane frequency. *Nature Geoscience*, 3:846–849.

Vickery, P. J., Skerlj, P. F., and Twisdale, L. A. (2000). Simulation of Hurricane Risk in the U.S. Using Empirical Track Model. *Journal of Structural Engineering*, pages 1222–1237.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Webster, P. J., Holland, G. J., Curry, J. A., and Chang, H. R. (2005). Changes in Trop-

ical Cyclone Number, Duration, and Intensity in a Warming Environment. *Science*, 309(5742):1844–1846.

Williams, C. and Seeger, M. (2001). Using the nystrm method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press.

Xie, L., Yan, T., Pietrafesa, L. J., Morrison, J. M., and Karl, T. (2005). Climatology and Interannual Variability of North Atlantic Hurricane Tracks. *Journal of Climate*, 18:5370–5381.

Zhang, P. (1996). Nonparametric importance sampling. *Journal of the American Statistical Association*, 91(435):1245–1253.

Appendix A

Notation Glossary

Table A.1: Notation Glossary

Notation	Description
Ω	Observed data on which the diffusion map is based
m	The cardinality of Ω
F	The (unknown) distribution of the data
G	The instrumental distribution
f	The density of F with respect to G
$\mathcal{X}$	Support of F
ℓ	(High) dimension of the observed data
Γ	Weighted graph $\Gamma = (\Omega, W)$
W	Weights for $\Gamma = (\Omega, W)$, where the weights connecting $x, y \in \Omega$ is given by $k_\epsilon(x, y)$
d	The application-specific, locally relevant distance measure
ϵ	The smoothing parameter of the diffusion map

Continued on next page

Table A.1 – continued from previous page

Notation	Description
k_ϵ	The smoothing kernel
p_1(x,y)	The one-step transition probability for a walk over Γ
p_t(x,y)	The t-step transition probability for a walk over Γ
T	The one-step transition matrix for a walk over Γ
λ_i	The eigenvalues of T
ϕ_i	The left eigenvectors of T
ψ_i	The right eigenvectors of T
$D_t(x, y)$	The diffusion distance
Ψ_t	The diffusion map
q	The dimensionality of the reduction
μ	The Lebesgue measure
S_ϵ	The stationary distribution of the Markov chain on the infinite state space $\mathcal{X}$
s_ϵ	The density of S_ϵ with respect to distribution the diffusion map is built on
A_ϵ	The diffusion operator – the continuous analog to T
$\psi_{\epsilon,i}$	The eigenfunctions of A_ϵ
$\mathcal{L}_d$	Proportion of inter-sample nearest neighbors
$\mathcal{R}_d$	The expected value of $\mathcal{L}_d$
h	The target function of the importance sampling
p^*	The density of the data distribution F with respect to the stationary distribution

Continued on next page

Table A.1 – continued from previous page

Notation	Description
s_i	Speed of a tropical cyclone at location i
θ_i	Direction of a tropical cyclone at location i
ν	The moment measure of the spatial point process
λ	The intensity function of the spatial point process