Adaptive Basis Density Estimation for High-Dimensional Data

Susan Buchman

January 14, 2010

Abstract

All high-dimensional density estimation techniques must make some assumptions about the underlying data distribution in order to be practical. In this proposal, I present work on a new method for high dimensional density estimation which assumes the ability to cheaply sample from an instrumental distribution which captures the low-dimensional structure in the data distribution. This assumption is satisfied in the application area of interest: modeling the distribution of tracks of tropical cyclones (TC) in the North Atlantic Ocean. Physical models are capable of generating realistic tracks, but not in the correct distribution over track space; my method allows for their use as instrumental distributions, anchoring the observed data in the vast high-dimensional space. Using orthogonal series density estimation with a basis that is adapted to the instrumental distribution, I produce a density for the data distribution with respect not to the Lebesgue measure, but with respect to the instrumental distribution, which has the potential to improve the rates of convergence of quantities of interest. Initial simulations support this hypothesis. I propose to extend this work to conditional density estimation to allow for the introduction of covariates, which when applied to the TC track data will reveal the relationship between spatial locations of TCs and climatic predictors. Furthermore, I will explore plug-in criteria for choosing optimal truncation points of the series, and for validating high-dimensional density estimates. I will establish consistency results for the procedures.

1 Introduction

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 and sampling from such estimates. Consider the paths of North Atlantic tropical cyclones (TC), some of which are shown in Figure 1(a). 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 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. 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.

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 applications presented above.

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. We present an approach which utilizes a *spectral connectivity analysis* (SCA) method (Lee and Wasserman, 2008) called diffusion maps. SCA reparameterizes 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 document, we present an high-dimensional density estimator and propose extending it to better address key questions regarding TC behavior, among other applications.

2 Basic problem and literature review

The low-frequency, high-severity nature of tropical cyclones (TC) 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 relationship between climatic predictors and the spatial variation of TCs, i.e. the TC tracks. 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 proposed for 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 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 parametric models, but the complex, non-linear nature of the data results in an explosion of parameters. We believe 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.

2.1 Data

As is standard in this field, we will rely on HURDAT, the "best track" database of North Atlantic tropical cyclones produced by NOAA Jarvinen et al. (1984). Of the 608 TCs between 1950 an 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.

2.2 Literature review

2.2.1 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. The National Oceanic and Atmospheric Administration characterizes them in the following way:

> 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. 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 them in order to make succinct claims about the past behavior of TCs on a much larger timescale, over which initial conditions are constantly changing.

The work that we pursue is a statistical model which can take advantage of dynamical models to inform the low-dimensional space in which tracks reside. We will not discuss the details of any dynamical models, although they may be used in our work as described below. 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 a similar 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 translation speed c and direction θ of a TC from time i to $i+1$:

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

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.

To the extent that the spatial distribution of TC tracks has been investigated, researchers have primarily focused on variability in landfall location. 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". 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 are really just an integral of the density 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.

2.2.2 Nonparametric high-dimensional density estimation

The alternative to the strong parametric assumptions of the previous models is to pursue *nonparametric density estimation.* All non-parametric 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 we are 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). 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 proportion of data points. So we very quickly find ourselves 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.

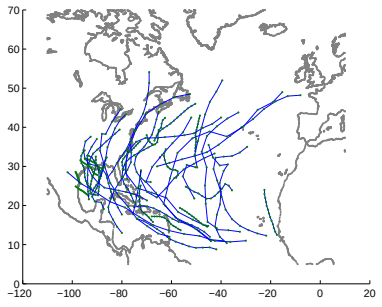
Due to the curse, kernel density estimation (Scott, 1992), a very popular technique, would not be appropriate for this application, as there were only 608 storms between 1950 and 2006.[1] As argued by Levina and Bickel (2004), "there is a consensus in the high-dimensional data analysis community that 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." We can only proceed if we make some reasonable assumptions about the data and provide 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 dimensions used to represent a datum. This is not typical of the data we consider here.

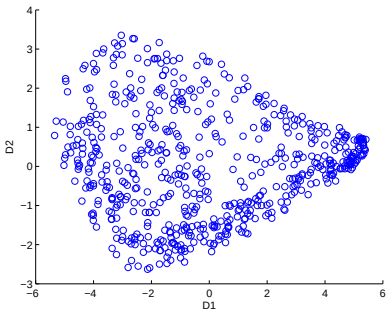
Previous work of ours in high-dimensional density estimation developed methodology that was appropriate for the nonlinear TC context (Buchman et al., 2009); it merely assumed that the data lie in a lower-dimension space. It involved constructing $\Psi_t : x \mapsto \Psi_t(x)$, a diffusion mapping (Section 3.1) of the high-dimensional data into a lower-dimensional space; performing standard density estimation in that reduced space; drawing a sample from the density estimate; and inverting back to the input space. An overview of this method, as applied to TC tracks, is shown in Figure 1.

The limitation of this method is that inverting back to input space (i.e. track space) — shown as the transition from Figure 1(d) to Figure 1(e) — requires a search that is not necessarily convex and can not be performed well in high-dimensional space. Thus the transition to the work in this document, which takes advantage of existing work in the generation of realistic tracks.

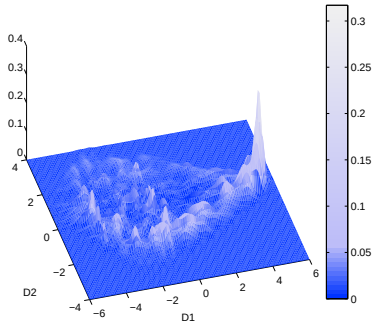
[1] It is standard not to perform estimation on years before 1950, as the lack of naval reconnaissance and, later, satellites have lead to undercounting and truncation in the database.



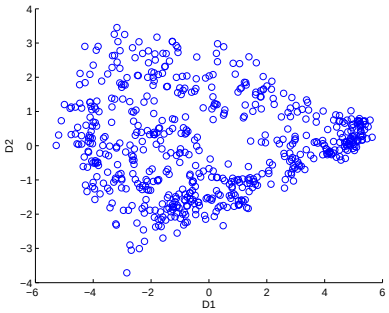
(a) A subset of the observed tracks.



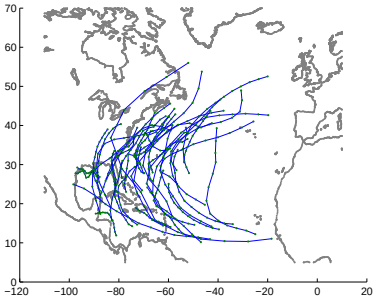
(b) The observed tracks in diffusion space.



(c) A density over diffusion space.



(d) A random sample from the density.



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

Figure 1: **An overview of our 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). An estimated density for the diffusion space data of (b) is visualized in (c), and a 608-element sample from that density is shown in (d). Each point in the sample can be interpreted as being associated with a new, as-yet-unobserved track. The sample is finally mapped back into track space; 40 randomly selected TCs of the sample are shown in (e).

2.2.3 High-dimensional model validation

Regardless of what form of high-dimensional density estimation is employed, generally it involves the selection of several model parameters — whether through cross-validation, maximum likelihood estimation, automatic "plug-in" methods, etc. — and we require a procedure for making a comprehensive evaluation of the resulting estimate.

Buchman et al. (2009) proposed a method which was appropriate for high-dimensional data and unspecific to any one density estimation method; instead, we produced a nonparametric high-dimensional verification technique that treats the particulars of the methodology as a black box. Our approach tests the hypothesis that two high-dimensional samples — the observed data and data simulated from the a density estimate — come from the same underlying distribution. 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. We make a connection between the choice of a 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

$$(X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n) \tag{1}$$

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

$$\lim_{n\to\infty} \mathcal{R}_d(\widehat{\mu}_n, \mu_X) = 0.5, \tag{2}$$

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.

To use the motivation behind the more formal notion of consistency with respect to the local distance metric to produce a test of our density estimate, we propose 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.

3 Preliminary methodological results

An overview of the method is as follows: the "curse of dimensionality" can be alleviated if we have a good, though not necessarily perfect, idea of where the density is located in the high-dimensional space — specifically, we assume that we can sample from an *instrumental distribution*, P, which puts most of its mass in the same region as the ℓ-dimensional *data distribution* P^*, although not in the same proportions. If we focus our efforts on targeting how P^* varies relative not with respect to Lebesgue measure μ on $\mathbb{R}^\ell$, but instead focus on a density with respect to the instrumental distribution, we believe that the convergence will exceed the worst-case $O(n^{\frac{-4}{4+l}})$. We do this by using *diffusion maps* to create an approximately orthnonormal basis with respect to the instrumental distribution, with which we can perform *orthogonal series density estimation*. However, instead of attempting to estimate the density $dP^*/d\mu$, we instead focus on dP^*/dP, as we will have more power to detect changes in P^* with respect to dP than $d\mu$. And the ability to sample from P means that the latter density can be used, via importance sampling, to answer most questions that a climate researcher might have.

Explicitly, the assumptions that we will be making are:

1. **"Good" data generation method.** We assume that we can generate inexpensively from an *instrumental distribution* P, where, at a minimum, $\mu \gg P \gg P^*$. Determining the conditions under which an instrumental distribution is good enough is one of the aims of this thesis.

2. **Inherently low-dimensional data.** We assume that the data distribution P^* takes a form that has a smooth (relative to P) and lower-dimensional (relative to ℓ) representation.

3.1 Introduction to diffusion maps

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

$$k_\epsilon(x, y) = \frac{1}{(4\pi\epsilon)^{l/2}} \exp(-d^2(x, y)/2\epsilon), \tag{3}$$

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 eigenvalues $|\lambda_0| \geq |\lambda_1| \geq \ldots |\lambda_{m-1}|$ 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*

$$D_t(x, y) = ||p_t(x, \cdot) - p_t(y, \cdot)||^2_{1/\phi_0} = \sum_{z \in \Omega} \frac{(p_t(x, z) - p_t(y, z))^2}{\phi_0(z)} \tag{4}$$

has an advantage over other common metrics for distances between distributions in that, in combination with the biorthogonal spectral decomposition of P^t:

$$p_t(x, y) = \sum_{j=0}^{m-1} \lambda_j^t \psi_j(x) \phi_j(y), \tag{5}$$

the diffusion distance can be approximated as

$$D_t(x, y) = \sum_{j=1}^{m-1} \lambda_j^{2t} (\psi_j(x) - \psi_j(y))^2 \tag{6}$$

and because of the decreasing eigenvalues, one can truncate at q terms for a suitably chosen q:

$$D_t(x, y) \approx \sum_{j=1}^{q} \lambda_j^{2t} (\psi_j(x) - \psi_j(y))^2. \tag{7}$$

Lastly, the *diffusion map*, defined as

$$\Psi_t : x \mapsto \begin{pmatrix} \lambda_1^t \psi_1(x) \\ \lambda_2^t \psi_2(x) \\ \vdots \\ \lambda_q^t \psi_q(x) \end{pmatrix}, \tag{8}$$

tidily reduces to

$$D_t(x, y) \approx \sum_{j=1}^{q} \lambda_j^{2t} (\psi_j(x) - \psi_j(y))^2 = ||\Psi_t(x) - \Psi_t(y)||^2. \tag{9}$$

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.

For our work, however, we are less interested in t and the full diffusion map, and primarily interested in the basis formed by the left eigenvectors of T and the connection to the stationary distribution of the random walk over G.

3.2 Connection to operators

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

$$\Omega_\epsilon(x, A) = \mathbb{P}(x \mapsto A) = \frac{\int_A k_\epsilon(x, y) dP(y)}{\int k_\epsilon(x, y) dP(y)}. \tag{10}$$

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 $dP/d\mu$. Let $p_\epsilon(x) = \int k_\epsilon(x,y)dP(y)$; the stationary distribution of the chain S_ϵ is

$$S_\epsilon(A) = \frac{\int_A p_\epsilon(x)dP(x)}{\int p_\epsilon(x)dP(x)} \tag{11}$$

and its density with respect to P is

$$s_\epsilon(x) = \frac{p_\epsilon(x)}{\int p_\epsilon(y)dP(y)}. \tag{12}$$

The diffusion operator A_ϵ is the continuous analog to the transition matrix T. Defined as

$$A_\epsilon f(x) = \frac{\int k_\epsilon(x,y)f(y)dP(y)}{\int k_\epsilon(x,y)dP(y)}, \tag{13}$$

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

$$\int \psi^2_{\epsilon,i}(x)dS_\epsilon(x) = \int \psi^2_{\epsilon,i}(x)s_\epsilon(x)dP(x) = 1 \text{ and } \int \psi_{\epsilon,i}(x)\psi_{\epsilon,j}(x)dS_\epsilon(x) = 0. \tag{14}$$

Thus the eigenfunctions are an orthonormal basis with respect to S_ϵ, the stationary distribution for the Markov chain with smoothing parameter ϵ. Our goal is to use the eigenvectors of T to estimate the eigenfunctions of A_ϵ and create an approximately orthonormal basis with respect to stationary distribution S_ϵ.

3.3 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 P^*. Suppose there exists a distribution P which puts most of its mass in the same region as the ℓ-dimensional *data distribution* P^*, although not in the same proportions. There need not be a known density for P — to return to the examples of Section 2.2.1, it might be a dynamical model. If one could sample inexpensively from this distribution *and* if one could estimate dP^*/dP, then most questions that could be answered with the more traditional $dP^*/d\mu$, where μ is the Lebesgue distribution, could be answered via rejection and importance sampling.

Our construction of dP^*/dP is built up as follow:

$$\boxed{f = \frac{dP^*}{dP} = \frac{dP^*}{dS_\epsilon} \cdot \frac{dS_\epsilon}{dP} = p^* \cdot s_\epsilon,} \tag{15}$$

where $p^* = \frac{dP^*}{dS_\epsilon}$ and, as before, $s_\epsilon = \frac{dS_\epsilon}{dP}$. The first factor of the right-hand side of Equation 15 is estimated using orthonormal series estimation, where the series in question is the estimated eigenfunctions of A_ϵ. 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 P cheaply, we can sample as many data points as we would like and improve the accuracy of the density estimate.

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, there is no natural way to take advantage of the subject matter expertise inherent in existing models and the realistic tracks that they generate.

3.4 Details of method

We start with an n-member sample Ω_S, the members of which are assumed to be an iid sample from P^*. We begin by generating Ω_I, a large sample of size m from the instrumental distribution P. (A notation guide for the various measures, sets, bases, etc., is available in Section 6.) For an appropriately chosen distance function d and a particular ϵ (in Section 3.7 we will discuss how to select ϵ, but for now we can treat it as fixed), we construct the $m \times m$ transition matrix T and decompose it into its eigenvalues and eigenvectors, as detailed in Section 3.1. The left eigenvectors $\{\psi_i\}$ will form the foundation of our estimation procedure. These eigenvectors can be rewritten as $\psi_i = \widehat{\psi}_{\epsilon,i}$, as they are estimates of the eigenfunction at our m observations from P. However, we will require the ability to estimate $\psi_{\epsilon,i}$ at all points in $\mathcal{X}$, not just the observed ones.

To do this, 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

$$\lambda_{\epsilon,i}\psi_{\epsilon,i} = A_\epsilon \psi_{\epsilon,i}. \tag{16}$$

holds. If we substitute Equation 13 in Equation 16, we see that

$$\psi_{\epsilon,i} = \frac{A_\epsilon \psi_{\epsilon,i}}{\lambda_{\epsilon,i}} = \frac{\int k_\epsilon(x,y)\psi_{\epsilon,i}(y)dP(y)}{\lambda_{\epsilon,i}\int k_\epsilon(x,y)dP(y)}, \tag{17}$$

which justifies the estimator

$$\widehat{\psi}_{\epsilon,i}(x) = \frac{\sum_{j=1}^n k_\epsilon(x,X_j)\psi_{\epsilon,i}(X_j)}{\lambda_{\epsilon,i}\sum_{j=1}^n k_\epsilon(x,X_j)}. \tag{18}$$

3.5 Estimating s_ϵ

How do we go about estimating $s_\epsilon = dS_\epsilon/dP$? Recall that

$$s_\epsilon(x) = \frac{p_\epsilon(x)}{\int p_\epsilon(y)dP(y)} \tag{19}$$

and

$$p_\epsilon(x) = \int k_\epsilon(x,y)dP(y). \tag{20}$$

Thus it is obvious that we can estimate p_ϵ as

$$\widehat{p}_\epsilon(x) = \frac{1}{q}\sum_{i=1}^q k_\epsilon(x,X_i). \tag{21}$$

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

$$\int p_\epsilon(y)dP(y) \approx \int \left(\frac{1}{q} \sum_{i=1}^{q} k_\epsilon(y, X_i) \right) dP(y). \tag{22}$$

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 P, call it S', of size m', and taking the second sample mean over this new data set:

$$\int p_\epsilon(y)dP(y) \approx \frac{1}{m'} \sum_{j=1}^{m'} \frac{1}{q} \sum_{i=1}^{q} k_\epsilon(X'_j, X_i). \tag{23}$$

Thus

$$\widehat{s}_\epsilon(x) = \frac{\frac{1}{q}\sum_{i=1}^{q} k_\epsilon(x, X_i)}{\frac{1}{m'}\sum_{j=1}^{m'} \frac{1}{q}\sum_{i=1}^{q} k_\epsilon(X'_j, X_i)} \tag{24}$$

3.6 Estimating p^*

In Section 3.4, we described how to estimate $\widehat{\psi}_{\epsilon,i}$, which approximates an orthonormal basis with respect to S_ϵ. Using orthogonal series density estimation (Crain, 1973; Girolami, 2002; Efromovich, 1999; Kronmal and Tarter, 1968; Diggle and Hall, 1986), we now estimate p^* using that basis:

$$p^*(x) = \sum_{i=0}^{\infty} \alpha_i \psi_i(x). \tag{25}$$

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

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

suggesting the unbiased estimator

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n}\psi_j(X_i). \tag{26}$$

In practice, we of course do not use an infinite basis as in Equation 25 — selection of the q-term truncation point is discussed in the next section, and will be a function of sample size, the geometry of P^*, and the smoothness of P^* relative to P.

So

$$\widehat{p}^*(x) = \sum_{i=0}^{q}\widehat{\alpha}_i\widehat{\psi}_{\epsilon,i}(x). \tag{27}$$

3.7 Selection of ϵ and q

We have, up until this point, treated ϵ and q as fixed, 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 it is *decreasing* q that increases bias but reduces variance.

There is work devoted to plug-in methods for selection q in the orthogonal series estimation literature, but it generally assumes a fixed basis. 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 18 and consider the effect that ϵ has on the estimators.

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

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

As in conventional MISE calculations, the second integral $\mathbb{E}\left(\int\widehat{f}(y)dP^*(y)\right)$ 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 first integral $\mathbb{E}\left(\int\widehat{f}^2(y)dP(y)\right)$ is an expected value, too, with respect to the instrumental distribution and can also be estimated by a sample mean. In other words, using Ω_Z, a sample from P of size m_{CV},

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

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

3.8 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 $X_i \sim P$, $E_{P^*}(h(x)) = \int h(x)\frac{dP^*(x)}{dP(x)}dP(x) \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i)f(X_i) \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i)\widehat{f}(X_i)$.

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 P, 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 $c_0 = \sup_x \widehat{f}(x)$,

$$c_0 \leq \left(\sum_{i=0}^{q} \frac{|\widehat{\alpha}_i|}{\lambda_{\epsilon,i}} max_{j:\mathtt{sgn}(\psi_i(X_j))=\mathtt{sgn}(\widehat{\alpha}_i)} |\psi_i(X_j)| \right) = c. \tag{28}$$

Rejection sampling can be performed using c, but whether it is efficient, i.e. whether this is a sufficiently tight bound to keep from rejecting all draws from the instrumental distribution, remains as proposed work (Section 5.3).

4 Preliminary applied results

We present an example of the power of the method of Section 3 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 2. Their underlying distribution P^* 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 have instrumental distribution P, 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 3, colored by their true density with respect to μ.

Given $n = 20$ and the two-dimensional data, I selected $m = 4000$ and $d(x,y) = ||x - y||$, and using the methodology of Section 3, $\epsilon = .35$ and $q = 4$ were chosen. Figure 4 shows the true versus estimated $\frac{dP^*}{dP}$, and we see that the method performs well, even in regions where dP^* is greater than 5 times dP. Table 1 compares the probability of being a specified region for four distributions: the true P^*; P^* as estimated by $\widehat{f}$; P^* as estimated by kernel density estimation; and P. (The two regions can be seen relative to the observed data in Figure 2; one region does not contain any data points.) The last one is provided to make clear that the method is not swamped by P; in fact, we can see from the first line of the table that in the first region (denoted by the green rectangle in Figure 2) where the true probability under P is an order of magnitude less than that of P^*, 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.

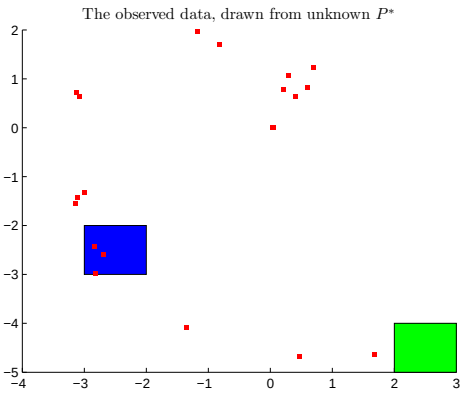


Figure 2: The 20 observed data points are shown as red squares. The green and blue rectangles correspond to sets on which different measures can be compared in Table 1.

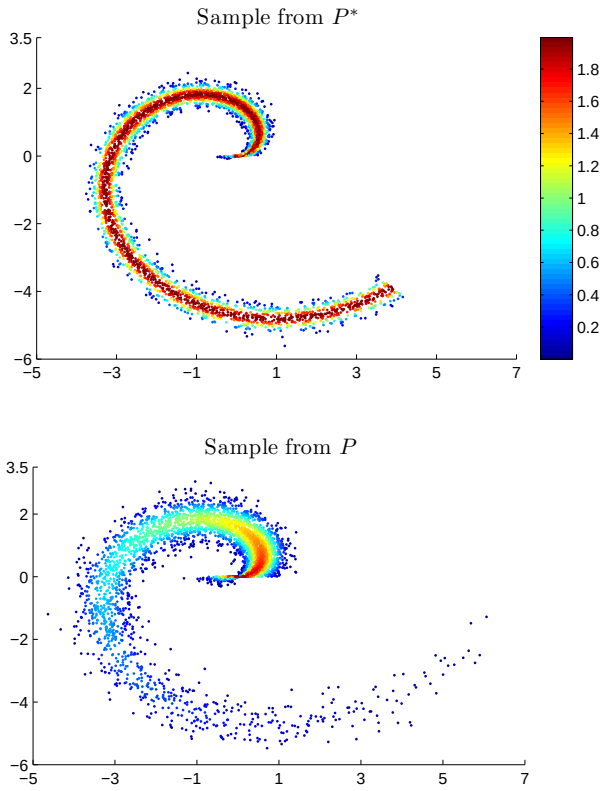


Figure 3: A sample of 4000 points from the data distribution P^* (top) and the instrumental distribution P.

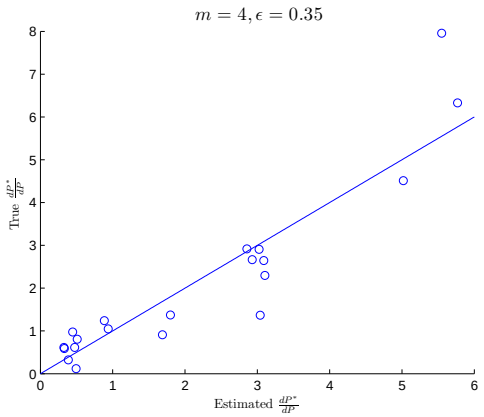


Figure 4: A plot of the true versus estimated values for dP^*/dP on the observed data Ω_S. The closer a point is to the line $y = x$, the better the density estimate at that point.

	P^*	P	$\widehat{f}$	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 1: The table compares the true measure of a set under both P^* and P to the estimates using adaptive basis estimation and kernel density estimation.

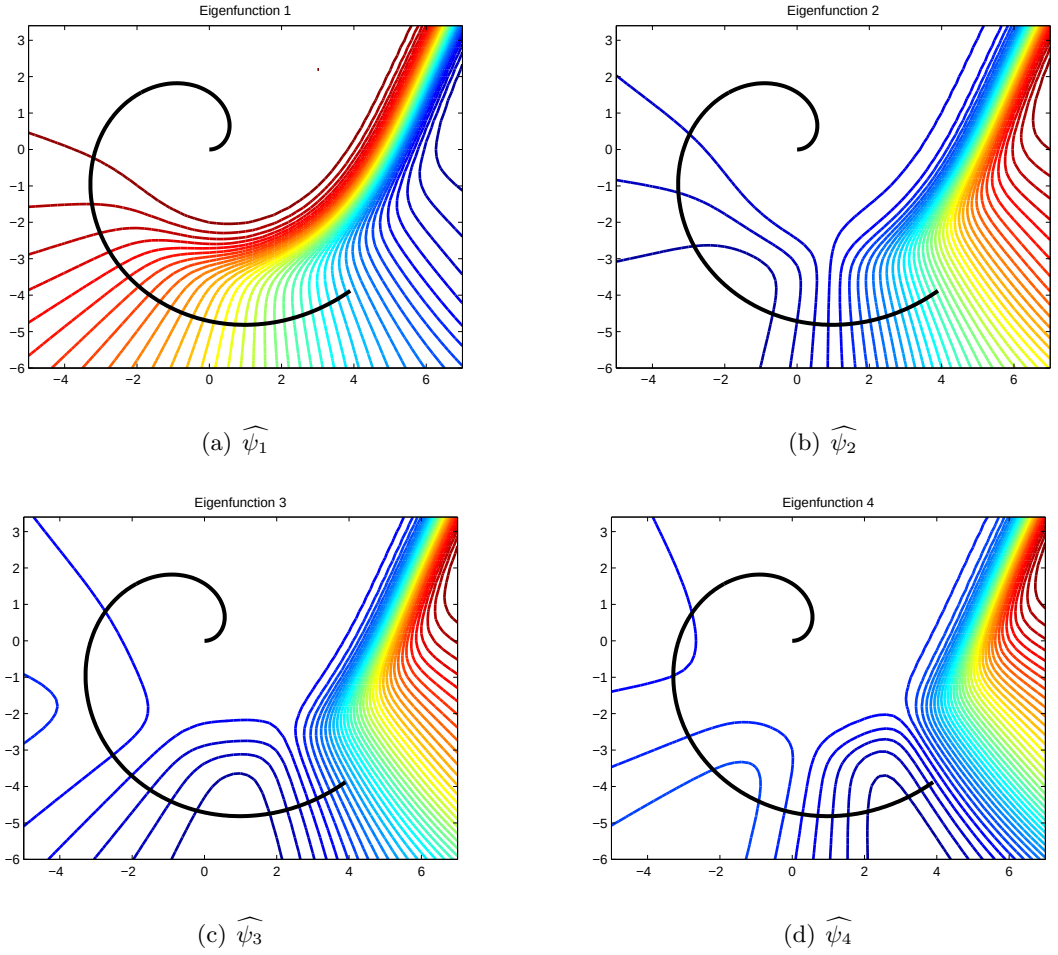
The estimated eigenfunctions, shown in Figure 4, really illuminate why the process is working. The zeroth eigenfunction is (necessarily) constant, and therefore omitted, but the first through fourth contours for the estimated eigenfunctions are quite intuitive. (The legends are omitted as the magnitude is not important for intuition; as α_i was negative for all $i > 0$, the contour coloring corresponds to blue being most dense, red being least dense.) The contours have been overlaid with the spiral corresponding to the mean curve of P^*.

The first eigenvector has contours *perpendicular* to the mean curve, and is essentially marking a point's mean location. The second, third, and fourth eigenvectors all descend steeply at the end of the mean spiral curve, which models the abrupt end to dP^*/dP 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.

5 Proposed work

The proposed work has both statistical and scientific aims. The statistical goals are to extend the methodology of Section 3 to *conditional* density estimation; 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 upon validation methods for high-dimensional data.

The scientific goal is to select from the suite of very detailed TC track models to serve as an



(a) $\widehat{\psi_1}$ (b) $\widehat{\psi_2}$

(c) $\widehat{\psi_3}$ (d) $\widehat{\psi_4}$

Figure 5: The estimated eigenfunctions are smooth relative to P, 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.

instrumental distribution and then work with the real TC track data and climatic predictors to study the effect of various covariates on the HURDAT data.

5.1 Incorporate covariates

A major goal of this thesis is to extend to adaptive basis density estimation to the realm of conditional density estimation (CDE). The majority of work on methods for nonparametric CDE has focused on the conditional kernel density smoother, in which the conditional density is estimated as the ratio of the kernel density estimates for the joint density of the response and the predictors and the marginal density of the predictors (Holmes et al., 2007; Gooijer and Zerom, 2003; Hyndman et al., 1996; Bashtannyk and Hyndman, 2001; Hall et al., 2004); however, that form of CDE will suffer from the same problems with high dimensional data as in the unconditional case.

Instead, I propose to adapt the method of Section 3. In the conditional case, for dependent

variable x and independent variable y, the form of the density changes from Equation 15 to

$$f(x|y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\theta_{i,j}\varphi_{i,j}(x,y). \tag{29}$$

I will begin by treating $\varphi_{i,j}(x,y)$ as a tensor-product basis bifurcated by the predictor and response — $\varphi_{i,j}(x,y) = \psi_i(x)\lambda_j(y)$. Under this formulation, Equation 29 becomes an example of a *varying coefficient model* (Hastie and Tibshirani, 1993):

$$f(x|y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\theta_{i,j}\varphi_{i,j}(x,y) = \sum_{i=0}^{\infty}\left(\sum_{j=0}^{\infty}\theta_{i,j}\lambda_j(y)\right)\psi_i(x). \tag{30}$$

Although varying coefficient models have primarily been applied to regression, approaching the density estimation in this way might lead to better estimation methods.

5.2 Work with real TC track data and 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.

5.2.1 Which climatic predictors?

An area of great concern is the effect that rising sea surface temperatures (SST) might have on the frequency and/or intensity of TCs. The density estimation method introduced in this paper can be applied to the question of how changes in sea surface temperature might affect the distribution of TCs. To do this, we might have to transform the covariate data, which may take values not over tracks but over the ocean. For example, SST varies both spatially and temporally. To assign a single SST value to a particular track, we may average SST over the track, attributing to a point on the track the SST at that location at the time that the track was actually there.

The potential of a model incorporating track-level SST data is illustrated in Figure 6. In this example, a three-dimensional diffusion map is created using 1000 TC tracks since 1900; this map is shown in the upper-left panel of the figure. Consider a point in three-dimensional diffusion space, $\mathbf{z}_0 = (0.39, 0.086, -0.0098)$. The upper-right panel of Figure 6 shows all of the tracks which are within a small diffusion distance (i.e. small Euclidean distance in diffusion space) of this point; these are a cluster of storms which remain far from the Atlantic coast. The dashed line in the lower-left panel shows the change in the density of tracks near $\mathbf{z}_0$ over all of the years, smoothed over time.

This panel also depicts the sea surface temperature at latitude/longitude (30W,15N)[2], chosen because it is close to the genesis point of the storms shown in the upper-right panel of Figure 6. The two vertical lines correspond to important time points in the improvement of storm observations: first in 1945 when plane-based observations began, and second in 1966 when satellite-based tracking began (Vecchi and Knutson, 2008). It is evident that once the improved data from satellites became

[2]From the Smith-Reynolds Extended Reconstructed Sea Surface Temperature Data Set (Smith et al., 2008).

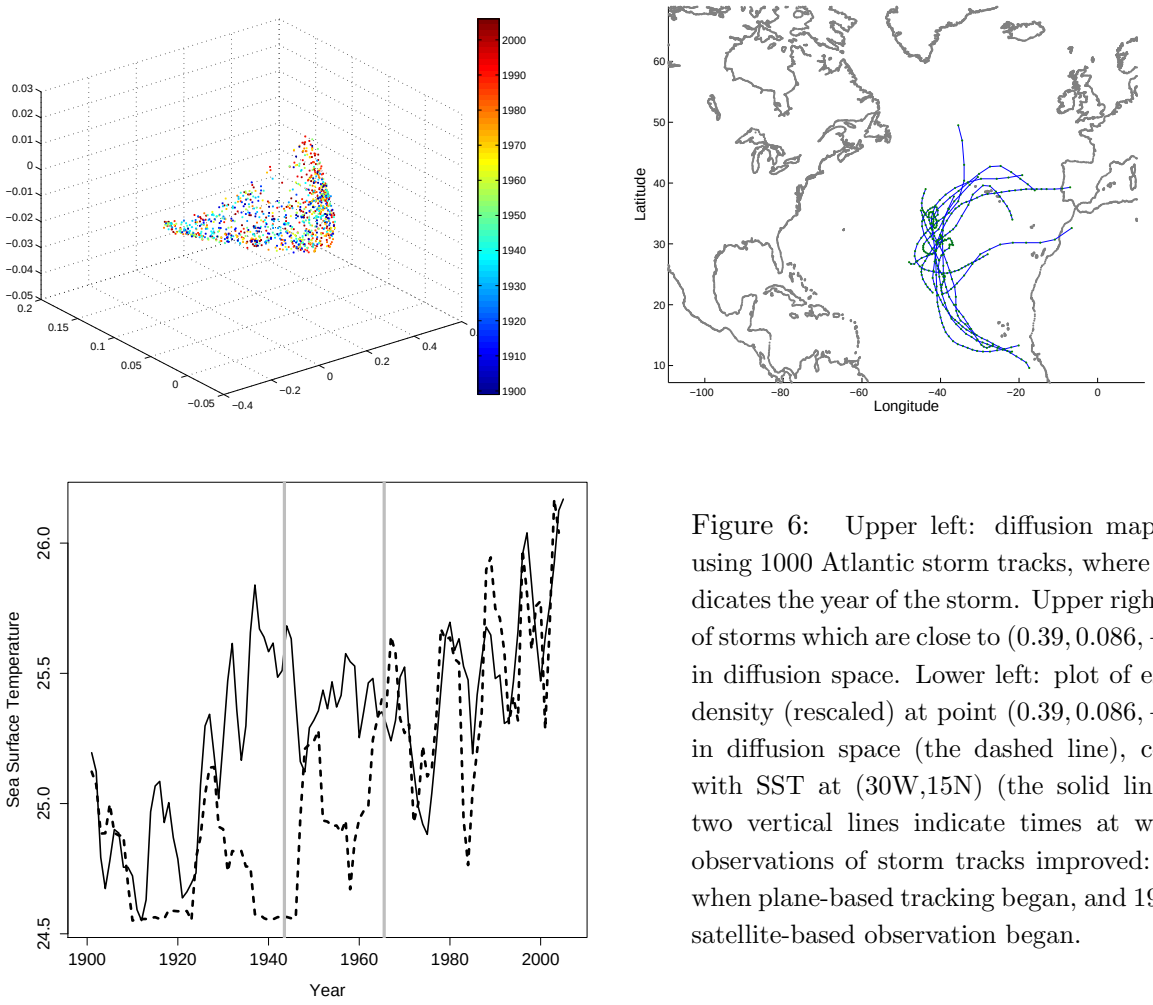


Figure 6: Upper left: diffusion map created using 1000 Atlantic storm tracks, where color indicates the year of the storm. Upper right: tracks of storms which are close to $(0.39, 0.086, -0.0098)$ in diffusion space. Lower left: plot of estimated density (rescaled) at point $(0.39, 0.086, -0.0098)$ in diffusion space (the dashed line), compared with SST at (30W,15N) (the solid line). The two vertical lines indicate times at which the observations of storm tracks improved: in 1945 when plane-based tracking began, and 1966 when satellite-based observation began.

available, there is a close correspondence between SST and storm occurrence. We plan to exploit these, and more sophisticated, temporal and spatial relationships.

In particular, our formal models for CDE will incorporate SST and other climate predictor variables such as the El Niño-Southern Oscillation and the North Atlantic Oscillation.

5.2.2 Selecting an instrumental model

To apply adaptive basis density estimation to the TC track data, we will need to select an instrumental distribution. Indeed, one reason that this model is so well suited to TC track modeling is the large body of work in creating TC track models that we can take advantage of. These models perform very well at generating realistic tracks, even though we suspect the do not generate them in the right proportions.

The National Weather Service's National Hurricane Service employs a large ensemble of models, with differing levels of transparency and ease of replication (NOAA, 2009). All the models of Section 2.2.1 could also serve as instrumental models. The final choice of instrumental model(s)

will primarily be driven by how feasible its replication is — is it well-documented? Is there existing code available? Etc.

5.3 Underlying assumptions and asymptotics

I propose to formalize this work by understanding the assumptions that the method is using implicitly, and by establishing asymptotic results.

The asymptotic work will revolve around three main questions: what, exactly, does it mean for an instrumental distribution P 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?

5.3.1 Sufficient properties of P

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 26, 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 P can improve those asymptotic results, but our method is, for now, heuristically motivated. We only require that P be a dominating measure for P^* and that P be "close enough" to P^*. Probing what it means to be "close enough", or the ramifications of choosing a poor P, is another aim of this thesis.

For example, in Equation 28, 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 P^* is with respect to P, 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 dP^*/dP, or by the second derivative's L_2-norm.

5.3.2 Approximately orthonormal basis

Even having chosen a sufficient P, the basis used is only approximately orthonormal with respect to S_ϵ; it is crafted from a very large, but finite, sample from P, 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 P and of the instrumental sample size used to construct the approximate and requires further investigation.

5.3.3 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 P.

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.

5.4 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 $\epsilon_1 < \epsilon_0$, $q^*_{\epsilon_1} \geq q^*_{\epsilon_0}$?

5.5 Adapt plug-in methods to account for ϵ

As discussed in Section 3.7, 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

$$\frac{2}{(n+1)} \frac{1}{n} \sum_{i=1}^{n} \phi_j^2(X_i) > \left(\frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) \right)^2 . \tag{31}$$

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.

We will 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 5.4 could aid in producing even more sophisticated plug-in methods.

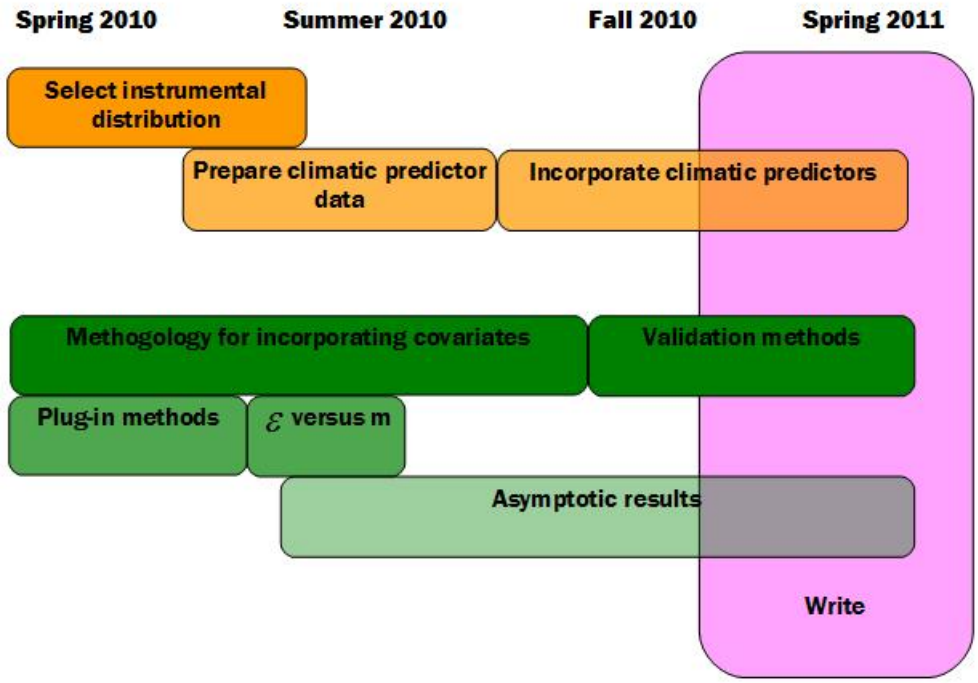
5.6 Validation methods

The validation technique in Section 2.2.3 is currently heuristically motivated. I aim 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 Notation Guide

Symbol	Meaning
P^*	The data distribution, i.e. the distribution we are trying to estimate
P	The instrumental distribution
S_ϵ	The stochastic distribution of the Markov chain
p^*	The density dP^*/dS_ϵ
s_ϵ	The density dS_ϵ/dP
f	The density dP^*/dP

7 Timeline



References

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.

Bashtannyk, D. M. and Hyndman, R. J. (2001), "Bandwidth selection for kernel conditional density estimation," *Computational Statistics and Data Analysis*, 36, 279–298.

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

— (2004), "Semi-supervised learning on Riemannian manifolds," in *Machine Learning*, pp. 209–239.

Buchman, S. M., Lee, A. B., and Schafer, C. M. (2009), "High-dimensional density estimation via SCA: An example in the modelling of hurricane tracks," *Statistical Methodology*, In Press, Corrected Proof, –.

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, 7426–7431.

Crain, B. R. (1973), "A Note on Density Estimation Using Orthogonal Expansions," *Journal of the American Statistical Association*, 68, 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, 230–233.

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

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

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

Gooijer, J. G. D. and Zerom, D. (2003), "On Conditional Density Estimation," *Statistica Neerlandica*, 57, 159–176.

Hall, P. (1986), "On the Rate of Convergence of Orthogonal Series Density Estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 48, 115–122.

Hall, P., Racine, J., and Li, Q. (2004), "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hall, P. and Tajvidi, N. (2002), "Permutation tests for equality of distributions in high-dimensional settings," *Biometrika*, 89, 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*, 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, 95–116.

Hastie, T. and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, 757–796.

Holmes, M. P., Gray, A. G., and Isbell, Jr., C. L. (2007), "Fast Nonparametric Conditional Density Estimation," in *The Learning Workshop (SNOWBIRD)*.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996), "Estimating and Visualizing Conditional Densities," *Journal of Computational and Graphical Statistics*, 5, 315–336.

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.

Kronmal, R. and Tarter, M. (1968), "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods," *Journal of the American Statistical Association*, 63, 925–952.

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, 1393–1403.

Lee, A. B. and Wasserman, L. (2008), "Spectral Connectivity Analysis," Submitted.

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 *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, eds. Meila, M. and Shen, X.

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*, MIT Press, pp. 849–856.

NOAA (2009), "Technical Summary of the National Hurricane Center Track and Intensity Models," `http://www.nhc.noaa.gov/modelsummary.shtml`.

Richards, J. W., Freeman, P. E., Lee, A. B., and Schafer, C. M. (2009), "Exploiting Low-Dimensional Structure in Astronomical Spectra," To appear in *Astrophysical Journal*.

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

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

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.

Smith, T. M., Reynolds, R. W., Peterson, T., and Lawrimore, J. (2008), "Improvements to NOAA's Historical Merged Land-Ocean Surface Temperature Analysis (1880-2006)," *Journal of Climate*, 21, 2283–2296.

Vecchi, G. and Knutson, T. R. (2008), "On Estimates of Historical North Atlantic Tropical Cyclone Activity," *Journal of Climate*, 21, 3580–3600.

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*, 1222–1237.

von Luxburg, U. (2007), "A Tutorial on Spectral Clustering," *Statistics and Computing*, 17, 395–416.

Williams, C. and Seeger, M. (2001), "Using the Nystrm Method to Speed Up Kernel Machines," in *Advances in Neural Information Processing Systems 13*, MIT Press, pp. 682–688.

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.