Research Statement

Tyrus Berry

The rapid expansion of model complexity and data availability in the applied sciences is quickly outpacing the classical approaches to mathematical modeling. The result has been a flood of new methods which draw from diverse research communities, ranging from computer science to biology. This emerging discipline, known as data science, has had such useful and successful applications that it has overwhelmed all attempts at developing a mathematical foundation. The limitations of data science are currently not well understood, and rigorous validation of algorithms is often impossible. Indeed, many of the achievements defy the conventional limitations of classical (parametric or simplistic nonparametric) modeling, such as the curse-of-dimensionality. We are now learning that these limitations can sometimes be overcome when data sets contain hidden mathematical structure (e.g. an embedded manifold) which is accessible to certain algorithms.

Whereas classical data analysis assumes a particular model, data science works by assuming (often implicitly, and sometimes unbeknownst to the data scientist) the existence of a hidden mathematical structure in the data. My work at the intersection of mathematics and data science is driven by my collaborations with scientists, such as biologists studying neuronal networks [14, 17, 19], physicists studying pattern formation in liquid crystals [11, 18], and meteorologists studying cloud formation and El Niño [6, 15]. The common challenge that I found in these applications was that the conventional ‘first principles’ modeling approach tends to break down for such phenomena, especially when interesting patterns exist. I view my role as finding ways to place these new techniques on a rigorous mathematical foundation. In fact, I think that mathematicians have a historic opportunity to fully leverage related work in geometry, topology and harmonic analysis to improve and refine data science methodology.

For example, consider the images in Fig. 1(a) which are frames from videos of a liquid crystal exhibiting complex spatiotemporal patterns. By applying our manifold learning methods [3, 4, 11] to the sub-images of the video we discovered the low-dimensional manifold shown in Fig. 1(b-d). A classical approach to analyzing this system would require coupled PDEs that depend on many unknown of parameters. Even if the model were known, it is difficult to derive to the low-dimensional manifold of solutions, but data science methods learn it directly from the data.

Figure 1: (a) Frames from a recording of liquid crystal experiment [11] at 10 V (top) and 12 V (bottom). (b) The 4-D geometry of the sub-image space defined by the liquid crystal dynamics at 10V; the fourth dimension is represented as color. (c) Sub-images drawn above their first two embedding coordinates which represents the phase of the rolls in the sub-image. (d) Higher dimensions represent roll width and angle.
The crucial breakthrough in the topological and geometric understanding of data was to assume that the data lies near an embedded manifold. This is a very broad assumption, since we do not specify a particular manifold, only the existence of a manifold. Once this assumption was made explicit, existing algorithms were shown to provably represent the manifold structure, but only under stringent conditions. Understanding these conditions mathematically has led to significantly improved algorithms, and some conditions have been alleviated entirely. For example, early methods inherited from the machine learning community were found to be consistent only when the data was sampled uniformly from the manifold, which motivated the diffusion maps algorithm that allows arbitrary sampling [21]. I have further extended these algorithms to allow manifolds with boundary [1] (see Fig. 2), non-compact manifolds [2, 3] (see Fig. 3), different geometries [4], and maps between manifolds [4, 5]. In applications, we have found that these new algorithms provide improved results (for example, better forecasts [6]) with smaller training data sets.

The key to moving data science forward is to place it on a firm mathematical foundation. This foundation should explicitly connect the algorithms of data science to appropriate assumptions using the mathematical languages of topology, geometry, and functional analysis. The hidden structure of data can then be rigorously connected to the algorithms that discover it, either by proving the efficacy of existing methods, or (more commonly) by modifying them to achieve such consistency. Truly understanding the diverse and evolving methods of data science requires establishing this mathematical foundation of hypotheses and provably effective methods. My overarching goal is to help build this mathematical foundation for data science. In particular, I aim to develop methods that can provably find, represent, and connect the patterns and structures hidden in data.

Figure 2: Estimating a density $f$ (left) on a manifold with boundary leads to bias in the standard estimator $f_{h,N}$. Our cut-estimator $f_{c,h,N}^-$ and higher-order cut-estimator $f_{c,2,h,N}^-$ correct the bias at the boundary [1].

Figure 3: (a) Data from a 2-dimensional Gaussian with a radial gap forming a non-compact manifold. (b-c) Standard graph-based manifold approximations do not represent the components of the manifold. (d) Density-weighted graph construction provably represents the manifold in the limit of large data [2].
1 Overview

A central motivation in my research is applications to dynamical systems, and my work has included applications to forecasting El Niño [6], pattern evolution in liquid crystals [11, 18], and evolution of neuronal networks [14, 17, 19]. I am also interested in applications to image/video [11] and document analysis. These applications involve high-dimensional data sets with high spatial and/or temporal resolution. Motivated by these challenges, my research interests can be divided into two main categories: the geometry of data, and its application to dynamical systems.

The central idea of the geometry of data is the assumption that a data set with Euclidean coordinates does not actually fill the entire space, but instead lies on a low-dimensional subspace which may have complex nonlinear structure. Often the assumption is that the data lies on or near a Riemannian manifold that is embedded in the ambient Euclidean space, and I call this assumption the geometric prior. If we can assume the geometric prior, then our goal is to learn the structure of the manifold from the data, represent this geometric structure numerically, and then use the structure to improve our analysis. For example, when the data is generated by a dynamical system, we can use the geometric prior to improve forecasting and even uncertainty quantification. One strength of the geometric prior is that the data requirements are independent of the dimension of the ambient space. However, the data requirements grow exponentially in the dimension of the manifold, and for complex problems the dimension of the manifold is often large. I am currently pursuing three methods of battling this curse-of-dimensionality, which are described in Section 3.

I have developed new theories and methods in manifold learning [1–5], state space reconstruction [6–11], ensemble Kalman filtering [7, 8, 12–14], and semiparametric statistical methods [15–17]. I envision these techniques as component parts of an emerging semiparametric approach to complex systems. In my view, classical approaches gravitate towards two different extremes. One extreme encompasses filtering techniques, which are fully parametric and thus require that an exact model is specified. On the other extreme are the non-parametric methods of state space reconstruction [26], which require no explicit model. In my collaboration with scientists, I have found that for many emerging problems neither of these extreme approaches is realistic. While a successful technique must fully utilize existing models and a priori structure, often these will not efficiently or adequately explain the observations. In these cases, we need to be able to adaptively quantify and correct modeling errors and discover residual structure in the data. In Section 2, I describe the nonparametric and semiparametric forecasting methods I have developed as well as the remaining challenges and current research directions.

2 The geometric prior for dynamical systems

In applying the geometric prior to dynamical systems, my primary focus has been on the forecasting problem, especially in cases where the models are unknown or existing models are subject to model error. In collaboration with John Harlim, my faculty mentor at Pennsylvania State University, I began by introducing a method for uncertainty quantification for gradient systems with an unknown potential [10]. Given a time series of data from the gradient system, we developed a method that provably represents both the state space and the unknown potential using the diffusion maps algorithm. A particularly challenging aspect of this work is that the standard diffusion maps algorithm [21] does not apply when the state space is not compact (for example if the state space is all of \( \mathbb{R}^n \)). To overcome this, we applied an idea from the kernel density estimation literature and extended the diffusion maps theory to variable bandwidth kernels [3], and we introduced the first method that provably represented non-compact manifolds.
In [10] we addressed three uncertainty quantification problems, namely filtering, forecasting, and response. In each case, we developed a fully nonlinear method that was provably correct or optimal in the limit of large training data. Our approach was based on the novel idea of using the diffusion maps algorithm to construct a basis for square integrable functions on the manifold described by the data. However, the restriction to gradient systems was an impediment to many interesting applications. To overcome this limitation, we introduced a completely new approach to representing stochastic dynamical systems based on training data [6].

In [6] we introduce the diffusion forecast algorithm, which approximates the solution semigroup to the Fokker-Planck PDE associated to an unknown stochastic system on an unknown manifold. In particular, we used the fact that the shift map (also known as the Koopman operator for deterministic systems) on a data set is an unbiased estimator of the solution semigroup. By representing the shift map in the basis of functions developed in [3,10], we find a Markov matrix that converges to the solution semigroup in the limit of large data. To perform the diffusion forecast, one can represent an initial probability density in the basis, apply the Markov matrix, and then reconstruct the forecast density.

In Fig. 4 we show the application of this method to data from a Lorenz-63 attractor. First, we learn the generalized Fourier basis using a variable bandwidth kernel and example eigenfunctions are shown in Fig. 4(a). Next we represent the shift map in this basis giving a Markov matrix shown in Fig. 4(b). Finally, we compare the model free forecast to an ensemble forecast using the true model. Having extended our nonparametric forecasting methodology to a very large class of stochastic dynamical systems on manifolds, we next addressed the restriction to low dimensional dynamical systems.

In order to apply our model-free forecasting algorithm to high-dimensional dynamics, we introduced a semiparametric framework in [16]. The key assumption is that an imperfect or incomplete parametric model is known, and that the model error component is low-dimensional. In [16] we used the diffusion forecast algorithm to correct the model error by learning a time dependent model for the existing parameters of the model. We also showed that it is possible to extract training data from noisy observation of the physical variables, and we developed a proof-of-concept semiparametric filtering approach that used noisy observations to determine the initial conditions for the diffusion forecast.
A remaining limitation of our semiparametric approach is that we assumed that the evolution of the unknown component of the model is independent of the existing model variables. In my research on model closure and reduced modeling [12], I showed that model error evolution is typically dependent on the existing variables, making this a severe limitation. My goal in this research direction is to address this problem by learning a model closure for the model error. In other words, I would like to extract a low-dimensional set of functionals on the existing variables that close the model error as a dynamical system. From a machine learning perspective, this is closely related to the feature identification problem, where the feature is the model error and the existing variables must be mined for information related to this feature. As a first step toward solving this feature identification problem, I developed the iterated diffusion map (IDM) in [5], which is described in Section 3.

3 The curse-of-dimensionality

The key limitation of nonparametric methods is that the amount of data required to learn a geometric structure grows exponentially as the intrinsic dimension increases. This curse-of-dimensionality is well known in kernel density estimation, where optimizing the bias-variance tradeoff leads to a bias that is proportional to $N^{-2/(4+d)}$ where $N$ is the amount of data and $d$ is the dimension of the manifold [1]. A similar law applies in manifold learning [4, 21], and for learning non-compact manifolds even obtaining this convergence requires variable bandwidth kernels [3]. Overcoming the curse-of-dimensionality is the key to wide applicability of manifold learning methods such as those described in Section 2. I am currently investigating three possible solutions to the curse-of-dimensionality in applications.

The first method that I am investigating is based on smoothness priors. This idea is motivated by the higher-order kernels that are used in kernel density estimation. These kernels (thought of as densities) have many zero moments, which results in a much smaller bias and allows the bias-variance tradeoff to be re-optimized with a reduced error. Of course, these higher-order kernels require additional smoothness assumptions on the density. The goal of this line of research is to develop higher-order kernels for manifold learning, which presumably will require additional smoothness assumptions on the manifold.

In fact, the diffusion map basis already applies a smoothness prior as shown in Fig. 5. The eigenfunctions of the Laplace-Beltrami operator successively minimize the roughness functional $\int |\nabla f| dvol$ and are therefore the smoothest functions on the manifold. By truncating functions and operators in this basis we are implicitly applying a smoothness prior. In Fig. 5 we consider
an important problem of trying to learn a function on the plane using only a few samples. Since
the function is very smooth, extension in the diffusion maps basis is vastly superior to using a
neural network which is based on a sigmoid basis. My goal is to enforce a smoothness prior on
the manifold, thereby allowing estimation of the basis on higher dimensional manifolds.

A second method of overcoming the curse-of-dimensionality is to assume that, although the
data lies on a high-dimensional manifold, there is a particular feature of interest which lies on a low-
dimensional submanifold. In [5] we developed the iterated diffusion map (IDM) which is a method
of representing the feature map via a geometric flow that contracts the irrelevant components of
the manifold described by the data. In the theory of local kernels [4] developed with my PhD
advisor Tim Sauer, we showed that it is possible to represent a diffeomorphism between two
manifolds using a linear map between appropriate diffusion coordinates. This theorem requires
using the more general class of local kernels to pull back the geometry of one manifold onto the
other via the diffeomorphism.

![Figure 6](https://via.placeholder.com/150)

Figure 6: Iterated diffusion map (IDM) on the annulus can extract the quotient manifold which the
feature of interest (indicated by color) depends on. Above: The feature of interest is the radius and the
IDM approximates a geometric flow which contracts the annulus onto a curve that represents the feature.
Notice that the IDM is capable of changing the topology of the manifold.

The idea of the IDM is to represent a feature map, which is typically not a diffeomorphism,
and the dimensions of the manifolds may not even be the same. In order to represent this map,
we develop a discrete approximation to a geometric flow, which contracts directions that are not
related to the feature. In Fig. 6 the discrete steps of this geometric flow are shown contracting
an annulus onto the feature defined by the radius. Once these unrelated directions have been
sufficiently contracted, we can then apply a local kernel to represent the final diffeomorphism to
the feature of interest. This provably represents the feature map in the case that the data set is
a product space made up of the feature of interest with another manifold of irrelevant variables.
By studying the associated geometric flow, I hope to extend this result to a larger class of feature
maps. Moreover, the IDM algorithm will be essential in extending the semiparametric approach
described in Section 2 to the case where the model error is not closed.

The third method of confronting the curse-of-dimensionality that I am researching is to assume
that much of the structure of the manifold is already known. In this case, the data should only be
used to learn the unknown component (which is assumed to be low dimensional). This approach
is motivated by semiparametric modeling in statistics, and led me to develop a semiparametric
framework for modeling dynamical systems [16]. In semiparametric forecasting, we assume that
an incomplete or imperfect model of a dynamical system is known, and we use a data set to
learn the missing or incorrect dynamics, thereby fixing the existing model, as described in Section
2. The power of this framework is that we do not make any assumptions on the form of the
model error or missing dynamics. Instead we build a fully nonparametric auxiliary model that
corrects/completes the existing model. Currently, I am continuing to research methods that can
leverage prior knowledge of the spatial structure of images to better understand the patterns in
the image content, and I describe these ideas in Section 4.
4 Meta-data and known structure in data

Existing methods for dimensionality reduction are powerful tools for finding hidden structure in data, however there are opportunities for improvement. Gains can be made by tailoring the approach to a known structure in the data, such as the pixel layout of an image or the ordering of a time series. For example, if we apply dimensionality reduction to a time series and we ignore the time-ordering, then often the dimensionality reduction will re-discover the time ordering and declare this to be the most important feature of the data. This is wasteful because we already knew the time ordering of the data. This shows how applying dimension reduction without regard to the known structure of the data will only be successful for relatively simple examples and will often rediscover features that were already known. In order to find the hidden structure of the data, we must first understand how to fully utilize the existing structure so that we do not waste time and data on rediscovering the known structure.

While it seems a daunting task to develop specialized techniques for every data type, many existing data structures are given by either a time ordering, or by meta-data describing a generalized spatial layout of the data coordinates. My next goal in this research program is to find and represent the natural geometry of data that has an existing spatial structure. For example, pixel coordinates are meta-data that form the spatial structure of an image. I propose using a data-adapted harmonic analysis in order to efficiently represent the spatial structure of the data. This data-adapted construction should start with the a priori spatial structure and then combine this with the geometry extracted from the data itself to form the data-adapted spatial geometry.

I am currently developing a two-step approach to utilizing the spatial structure of data. In the first step, a user-supplied spatial structure, such as a pixel layout, is used to develop a harmonic analysis on the supplied spatial structure. By representing the data in a generalized Fourier basis, constructed by applying diffusion maps to the meta-data, we can naturally leverage the known spatial structure to improve our understanding of the data. However, often the spatial structure that is available does not fully capture the correct notion of similarity in the data set. In order to incorporate the content of the image, we represent the image on an over complete set of sub-images that are translated and rotated sub-images taken from the image itself. The weakness of current techniques (such as optimal basis construction), is that they do not recognize the geometry of the sub-image space. By learning the geometry of the sub-image space, we can efficiently represent the image as a skew product between the known spatial geometry (represented by translations, rotations, and dilations) and the learned geometry of the sub-images.

The adapted spatial analysis allows us to decouple the spatial structure of the image, which is known a priori, from the hidden variables that describe the content of the image. In the future, I also plan to use generalized wavelet bases to achieve finer local control of the quotient space and to exploit symmetries in the existing spatial structure. The ability to isolate some features of data requires certain symmetries in the underlying geometry. The existence of redundant structures and related symmetries may imply the existence of hidden algebraic structures. The idea of invariance under the diffusion geometry as a generalization of isometry [23, 24] has been developed into a method for finding intrinsic symmetries in a data set using the eigenmodes of the Laplace-Beltrami operator with isolated eigenvalues [25]. However, isolated eigenvalues correspond to simple reflection symmetries, and I am more interested in repeated eigenvalues that correspond to continuous symmetries, represented by subgroups of orthogonal matrices. These complex group structures correspond to significant inefficiencies in the data representation. Moreover, in the case of many interconnected symmetries, there may be efficient representations that correspond to group factorizations. Finally, by finding partial symmetries it may be possible to produce new, unobserved data by generating the data points that would complete these symmetries.
5 References


