Research Statement

Tyrus Berry

November 12, 2012

1 Introduction

The rapid expansion of model complexity and data availability in the applied sciences is quickly outpacing the classical approaches to mathematical modeling. My goal is to develop methods which can provably find, reconstruct, and represent the hidden stochastic, geometric, and algebraic structure of data. I am motivated by applications such as image and signal analysis, and dynamical systems including dynamical networks; and currently I am collaborating with physicists studying pattern evolution in liquid crystals [2], and biologists studying the evolution of neuronal networks [3]. These applications involve high-dimensional data sets with high temporal resolution such as video and micro-electrode array data. Motivated by these challenges, I am working to extend the current methods of dimensionality reduction and discrete geometry such as Diffusion Maps and the Discrete Exterior Calculus and integrate these methods with classical techniques for data analysis.

My work so far has focused on the problems of data assimilation, prediction, and control for high dimensional dynamical systems. Thus I have worked with techniques such as state space reconstruction [2], ensemble Kalman filtering [1, 4], limiting behavior of difference equations [5], and semi-parametric statistical tests such as the Cox method [3]. I envision these techniques as component parts of an emerging semi-parametric 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, which require no explicit model. In my collaboration with scientists I have found that for many emerging problems neither of these extreme approaches to 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.

I am approaching this *semi-parametric* middle ground from both of the extremes. From the non-parametric extreme, in [2] I have shown how to find the intrinsic geometric structure of time series data. From the fully parametric extreme, in [1] I have shown that model error can be automatically traded off for system noise to achieve significant improvement in existing filtering techniques. Each of these theoretical results were directly motivated by the real world problems that my collaborators in applied science were confronting and have lead to significant improvements in their analyses. For example, in [2] we developed a new algorithm which provably reconstructs the intrinsic geometry for a dynamical system from an observed time series. In this case the a priori structure is that of a stationary temporal evolution, but more generally a priori structures could be given by spatial relationships between coordinates (such as the layout of pixels in an image) or even incomplete or approximate models. In order to discover and represent the hidden structure of data, I am currently working to extended methods for discrete geometry, such as Diffusion Maps, and integrate these methods with classical time series analysis and data assimilation techniques. These initial results suggest a more far-reaching program to develop a semi-parametric technique which adapts an approximate model to the data and then extends the model using new variables discovered through a non-parametric examination of the residuals.

In the following sections I will first address the weakness in current dimensionality reduction techniques and how I plan to address them. In Section 2, I show how dimensionality reduction techniques should use the existing structure of the data. I overview my work in [2] which shows how to use the temporal structure of data and I describe several extensions of this research which I am currently pursuing. Next I discuss new techniques which I am developing to isolate the existing spatial structure found in many data sets and I give examples which show how this makes finding the hidden structure of data more practical by reducing the data requirements. In Section 3 discuss my long term goal of extending the dimensionality reduction technique known as Diffusion Maps, which was developed by Coifman and Lafon in [8]. I believe diffusion maps can be extended to describe more general geometric structures and can facilitate the discovery of higher order geometric and topological features such as continuous symmetries and important quantities from Hodge theory. Finally, in Section 4 I describe how the improvements to dimensionality reduction will lead to a semi-parametric modeling technique by integration with my work in [2] and [1].

2 Utilizing the known structure of data

In my experience working with real world applications, I have found that while existing methods for dimensionality reduction are powerful tools for finding hidden structure in data, there are significant opportunities for improvement. One approach to improving dimension reduction is through a better theoretical understanding of the methods, and in the next section I propose extending existing methods to more general geometries and constructing higher order geometric operators. However, an often overlooked, and possibly more important, aspect of dimension reduction is tailoring the approach to the known structure of 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. However, this is incredibly wasteful because we already knew the time ordering of the data. Thus, simply 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 which 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. In this section I first discuss how our work in [2] recovers the natural geometry for a time series and I discuss some important extensions which will also help motivate the theoretical goals of the next section. 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 which give 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.

Diffusion Mapped Delay Coordinates (DMDC) [2] is the first step towards a practical approach to reconstructing and analyzing the intrinsic geometry of time series. First, DMDC uses a weighted time-delay embedding in the spirit of Takens' to reconstruct the latent state space. Importantly we prove that our embedding also reconstructs the latent geometry of the time-series and projects the dynamics onto the most stable component. Next, DMDC uses the Diffusion Maps dimensionality

reduction technique to find the best low dimensional set of variables which accurately represent the latent geometry. However, although DMDC is able to project onto the stable dynamics regardless of dimension, when the stable dynamics are high dimensional a further reduction may be desired. Moreover, while some applications only require the stable dynamics, control approaches such as OGY [13] and Pyragas [14] also require knowledge of the unstable dynamics. These extensions will require a complete reconstruction of the Lyapunov geometry for the dynamical system, a difficult problem which will require extending the reconstruction theorem of [2] as well as new methods for discrete representation of anisotropic geometries which are discussed in Section 3.

While my work with DMDC has focused primarily on the temporal structure of data, many data sets will also have a spatial structure, which should inform our analysis. The most obvious examples are images, and it is immediately clear that treating the pixels as independent and unrelated observations (as PCA does for example) is a considerable underutilization of the data. More generally, any meta-data which gives a notion of distance between data coordinates gives an a priori spatial structure. For example, a collection of survey data may contain demographic or location meta-data which can be used to define an a priori distance between the surveys. Currently, using meta-data requires *ad hoc* solutions, each specially tuned to various data types and applications, but the growing complexity of these problems calls for a more unified strategy.

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 bases, constructed by applying diffusion maps to the meta-data, we can naturally leverage the known spatial structure to improve our understanding of the data. For example, consider two images which are each all white except for a single black pixel. If we compare these images in the basis of pixels then their distance will not depend on the proximity of the black pixels, so the natural notion of distance is not captured in this basis. However, if we represent the images in a Fourier basis and compute the distance using low frequency components then when the pixels are close to each other the low-pass images will be similar and the distance between them will be smaller. This shows how a spatial Fourier basis can incorporate the natural distance of the spatial structure into our geometry. We refer to this first step as the *spatial analysis*. However, often the spatial structure that is available is only does not fully capture the correct notion of similarity in the data set.

Consider a video of an object moving in the plane of the image and simultaneously rotating about its center. The content of each image can be naturally represented in a three dimensional space given by the coordinates of the object and the angle of its rotation. The weakness of current techniques (such as optimal basis construction), is that they do not recognize the geometry of the sub-image space, and thus they cannot generalize to slightly different elements. To represent the rotating object the optimal basis will require a different basis element for each angle in the rotation. While dimensionality reduction techniques can find the three dimensional latent coordinates, for practical purposes this will require enormous amounts of data. This is because current dimensionality reduction techniques work by interpolation and thus they must have examples of almost every orientation at *each location*. Thus, the current dimensionality reduction techniques are re-discovering the spatial layout of the image at the same time as they are finding the hidden structure of the rotation variable.

The adapted spatial analysis allows us to decouple the spatial layout of the image, which is known a priori, from the hidden rotation variable. Our initial spatial analysis, by representing the images in the low frequency spatial Fourier modes, would capture the location of the object within the image, however this projection would blur the image so that the rotation is not captured. Thus the a priori spatial layout of the pixels reveals two of the three latent coordinates. To find the

remaining structure we can now focus on the high frequency spatial Fourier modes. Effectively this produces the quotient space where the location information is removed, and now the dimensionality reduction technique only needs to find the remaining one-dimensional structure of the rotation variable, rather than the full three dimensional structure of the data set. The key to the wide applicability of this quotient technique is to that the Fourier analysis can be adapted to any spatial structure, which allows us to decouple any known spatial structure from the hidden structure. 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.

3 Finding hidden structure in data

Overcoming current modeling challenges requires radically transforming the current state of the art of dimension reduction. My thesis research is focused on the development of new approaches to interpretation, resolution, and feature extraction for high-dimensional dynamical data. Diffusion maps, a technique developed by Coifman and Lafon in [8], is the first technique which provably reconstructs the geometry of a manifold from discrete samples in Euclidean space for an arbitrary positive sampling density. In this section I focus on overcoming critical weaknesses in the current diffusion map approach, by extending the technique to more general geometries which allow drift and anisotropy in the Laplace-Beltrami operator. Next I propose merging diffusion maps with the discrete exterior calculus, developed in [6, 7] which will allow the construction of higher order geometric operators. This will allow novel methods of using the diffusion geometry, to find symmetries in the adapted geometry that represent intrinsic features of the data.

As we show in [2], diffusion maps allows approximation of the second order diffusion term of an evolution. We propose to extend the approximation to match the first order drift term of the evolution and improve the second order term by approximating the local covariance structure of the diffusion. From a geometric point of view, instead of defining a symmetric diffusion metric, we will define an asymmetric generalization of the metric called a Finsler function and model our data as a Finslerian manifold (a generalization of a Riemannian manifold). This approach has a particularly valuable interpretation for dynamical data, where the operators we construct can evolve a distribution of events forward in time to predict their likelihood. The ultimate goal would be to find a procedure which uses the local covariance structure of the data (and the time ordering if available) to provably construct a discrete approximation to the Finsler function for our manifold. This is a natural generalization of diffusion maps, which provable constructs the heat kernel on a manifold, and thus our Finsler function should be given implicitly by asymmetric kernel which is discretely approximated for each pair of data points in a large sparse matrix.

To motivate the next natural extension of diffusion maps, imagine a video where one pendulum is placed behind the other, such a data set is intrinsically two-dimensional and cannot be decomposed using the existing structure of the data because the pendulums occupy overlapping spatial and temporal scales. This simple example illustrates the problem posed by intrinsically high-dimensional data, and in general the relationship between the variables could be arbitrarily complex. Diffusion maps finds a low dimensional embedding of this data set by approximating the eigenfunctions of Laplacian on the underlying geometry, however, for a high dimensional manifold this is insufficient to separate the independent components. In fact, we will require the full power of de Rham cohomology and Hodge theory to carry out this goal.

Fortunately, recent work [6, 7] has paved the way for de Rham cohomology to be computed efficiently in engineering applications using a discrete version of exterior calculus. In fact we will show that the de Rham cohomology is a natural generalization of diffusion maps to higher-dimensional structures. Our proposal is to complete the theory which connects the discrete exterior

calculus to the diffusion maps construction and then to develop practical algorithms that exploit the cohomology to extract features from high-dimensional structures in data.

The de Rham cohomology allows one to extend the notion of linear independence to curved geometric structures. Equivalence classes of cochains correspond to separate intrinsic features in the data that can be smoothly deformed to match one another. These equivalence classes can be identified by finding the harmonic forms of higher order Laplacian operators [11, 12], and we will see that in the context of the discrete exterior calculus this can be represented as a large eigenvector problem. In our example of the two overlapping pendulums, each image of the movie can be defined by two intrinsic angles which define a torus. The first de Rham cohomology group for the torus contains exactly two equivalence classes which correspond to the two angles that define the separate pendulums. Thus by projecting onto one of the equivalence classes we can extract the corresponding pendulum. This example illustrates how the de Rham cohomology captures the structure of a multidimensional manifold and can be used to decompose the manifold into its intrinsic topological features.

Finally, 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 [17, 18] 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 [19]. However, isolated eigenvalues correspond to simple reflection symmetries, and I am more interested in repeated eigenvalues which correspond 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 which 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.

4 Making the ensemble Kalman filter adaptive

Kalman filtering is a well-established part of the engineering canon for state and uncertainty quantification. In the case of a linear plant with Gaussian system and observation noise, Kalman's algorithm is provably optimal, provided that the exact model and noise statistics are known. Further uses of the Kalman filter have included parameter fitting, where the parameters are treated as states with trivial dynamics, which is typically called "dual estimation". Our use of EnKF for fitting parameters, including connection coefficients in networks [4], is a challenging problem for the ensemble Kalman filter as well as other known algorithms. Our pursuit of this goal, which pushes the efficiency of the Kalman filter to its limit, led us to search for ways to optimize the filtering process.

For nonlinear systems, the Extended Kalman Filter (EKF) and Ensemble Kalman Filter (EnKF) provide two different ways to make use of the Kalman update. While these techniques have considerable success in applications, there are theoretical issues which can lead to filter divergence or poor performance. The root of these issues is the assumption that the error of the current state estimate (called the background distribution) has a Gaussian distribution. While this is true for linear systems it fails in general for nonlinear systems, which implies that the Kalman update cannot correctly integrate the information of the current state estimate with the information of the incoming observations. This causes the covariance estimate to diverge from the true background covariance.

Various ad hoc solutions, such as covariance inflation, have been proposed. In [1] we replace

these *ad hoc* covariance inflation strategies with a theory-based, realtime approximation of the covariance matrices of the noise processes. Our technique is inspired by the innovation correlation method of Mehra [24, 25]; however, significant changes are required to lift this technique to the nonlinear domain. We show that when the correct model is given, the estimated covariance matrices approximate the true matrices. Moreover, when model error is present the system noise covariance automatically adjusts in a way that compensates for the model error. In both cases our adaptive filter leads to significant reductions in the RMSE of state estimates.

I plan to extend this foundational work in several ways which will increase the applicability of our adaptive EnKF. First, the method of Mehra requires full observability, and in cases of partial observability he bypass the noise covariance by using the stationarity of the optimal Kalman gain. Of course, for nonlinear systems the optimal gain is not stationary. To recover the system noise covariance in the case of partial observability we propose an augmented observation formed by concatenating several iterations of the dynamics, which can be compared to time-delayed observation vector. For a generic observation, the time-delayed observation vector should fully represent the underlying state space making the augmented observation invertible [26]. We believe that this augmented observation will not only solve the partial observability problem, but may also improve the stability of the Kalman filter by including more observed information in each Kalman update.

A second natural extension would be to allow non-additive system noise, which is usually accomplished in the EnKF by augmenting the state vector and covariance matrix to include the realization of the noise in the ensemble. A particularly interesting consequence is that non-additive noise could help compensate for the multiplicative effects of Lyapunov exponents in a strongly nonlinear system.

Finally, it may be helpful to allow the covariance of the system noise to be state dependent, for example by recording each estimated covariance and using a local interpolation at each filter step. This would allow for heteroscedastic statistics, but even more interesting is that it could effectively provide a local correction to the Kalman update, which may help compensate for the incorrect assumption that the state estimate error is Gaussian in the extended and ensemble Kalman filters. The bigger picture is that accounting locally for nonlinearity and resultant Lyapunov exponents suggests the possibility of an *optimal* filter for a nonlinear system which would revolutionize the theory and applications of filtering.

Our current adaptive EnKF algorithm can be considered a coarse approach to extending the model; we determine an additive system noise term which is sufficient to explain the observation errors. In [1] the implied system noise found by our algorithm greatly improves the state estimates even in cases of large model error. While this first step only uses an additive stochastic term to compensate for model error, it opens the door to more advanced analysis of the residuals. Another promising candidate for adapting a parametric filter to the residuals is to use our non-parametric algorithm, Diffusion Mapped Delay Coordinates (DMDC). For a Kalman filter with an incomplete or incorrect model, the residual errors should contain a predictable component. Applying DMDC to these residuals would produce an optimal set of new variable to append to the state space. This leads to several interesting and promising research questions. First, what dynamics should be used for these new state variables, and how do we modify the Kalman update to estimate and predict with these new variables? Second, can we now use these new variables to improve our existing models by allowing the existing parameters to depend on the new state variables? Finally, can we prove that analyzing these residuals can actually reconstruct the missing or erroneous parts of our existing model? Answering these questions will undoubtedly require further developments in the theory of DMDC and the dimensionality reduction techniques discussed above.

5 References

[1] T. Berry and T. Sauer, *Adaptive ensemble Kalman filtering of nonlinear systems*, Preprint (2012).

[2] T. Berry, T. Sauer, R. Cressman, and Z. Greguric-Ferencek, *Time-scale separation from diffusion-mapped delay coordinates*, Submitted to SIAM Journal on Applied Dynamical Systems. (2012).

[3] T. Berry, T. Sauer, F. Hamilton, and N. Peixoto, *Detecting Connectivity Changes in Neuronal Networks*, Journal of Neuroscience Methods 209 (2012), 388-397.

[4] T. Berry, T. Sauer, and F. Hamilton, *Unscented Kalman Filtering for Link Detection and Tracking in Neural Networks*, Preprint (2012).

[5] T. Berry and T. Sauer, *Convergence of periodically-forced rank-type equations*, Journal of Difference Equations and Applications 17 (2011).

[6] M. Desbrun and E. Kanso and Y. Tong, *Discrete differential forms for computational modeling*, SIGGRAPH '06 (2006), 39–54.

[7] Matthew Fisher, Peter Schrder, Mathieu Desbrun, and Hugues Hoppe, *Design of Tangent Vector Fields*, ACM Transactions on Graphics **27(3)** (2007).

[8] R. Coifman and S. Lafon, *Diffusion Maps*, Applied and Computational Harmonic Analysis **21** (2006), 5-30.

[9] V. Garcia and E. Debreuve and M. Barlaud, *Fast k nearest neighbor search using GPU*, CVPR Workshop on Computer Vision on GPU, 2008.

[10] D. Strook, *An Introduction to the Analysis of Paths on a Riemannian Manifold*, American Mathematical Society, 2000.

[11] J. Jost, *Riemannian Geometry and Geometric Analysis*, Springer-Verlag Berlin, Berlin, 2002.

[12] S. Rosenberg, *The Laplacian on a Riemannian manifold*, Cambridge University Press, New York, NY, New York, New York, 1997.

[13] E. Ott, C. Grebogi, and J. Yorke, *Controlling chaos*, Phys. Rev. Lett. **64** (1990), 11961199.

[14] K. Pyragas, *Control of chaos via an unstable delayed feedback controller*, Phys. Rev. Lett. **86** (2001), 2265 - 2268.

[15] T. J. Kaper, H.G. Kaper, and A. Zagaris, *Two perspectives on reduction of ordinary differential equations*, WILEY-VCH Verlags **278** (2005), 1629-1642.

[16] C. W. Gear, T. J. Kaper, I. G. Kevrekidis, and A. Zagaris, *Projecting to a Slow Manifold: Singularly Perturbed Systems and Legacy Codes*, SIAM Journal on Applied Dynamical Systems **4** (2005), 711-732.

[17] D. Raviv and M. Bronstein and G. Sapiro and A. Bronstein and R.Kimmel, *Diffusion symmetries of non-rigid shapes*, In Proc. 3DPVT, 2010.

[18] A. and Bronstein Bronstein M. and Kimmel, *A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching*, Int. J. Comput. Vision **89** (2010), no. 2-3, 266–286.

[19] M. and Sun Ovsjanikov J. and Guibas, *Global intrinsic symmetries of shapes*, Proceedings of the Symposium on Geometry Processing, 2008, pp. 1341–1348.

[20] L. Arnold, *Random Dynamical Systems*, Springer-Verlag New York, Inc., New York, New York, 1998.

[21] R. Coifman and M. Maggioni, *Diffusion wavelets*, Appl. Comp. Harmonic Anal. **21** (2006), no. 1, 53 - 94.

[22] A. and Maggioni Szlam M. and Coifman, *Regularization on Graphs with Function-adapted Diffusion Processes*, J. Mach. Learn. Res. **9** (2008), 1711–1739.

[23] S. Mallat, *A Wavelet Tour of Signal Processing*, 2008.

[24] R. Mehra, *On the identification of variances and adaptive Kalman filtering*, IEEE Trans. Auto. Cont. **15** (1970), 175-184.

[25] ———, *Approaches to adaptive filtering*, IEEE Trans. Auto. Cont. **17** (1972), 693-698.

[26] T. Sauer, J.A. Yorke, and M. Casdagli, *Embedology*, Journal of Statistical Physics **65** (1991), 579-616.