Manifold Learning to Enable Equation-Free Modeling of Chemical Engineering Systems

Proposed by David Sroczynski

under the supervision of

Professor Yannis Kevrekidis

11/04/2015

Contents

1 Introduction

Virtually every field of science and engineering uses modeling to expand and elucidate the information available from experiments. These models range from simple, one-variable ODEs for a CSTR to computational fluid dynamics code. Experimental data is generally noisy and incomplete; a well-developed model can bring clarity to complicated systems. Models have the additional advantage that they can be initialized as often as desired in whatever state is desired, while experiments can be challenging or expensive to initiate. The obvious caveats are that a model must be accurate enough to capture the "important" information while being compact enough to be solved in a reasonable timeframe.

The field of dimensionality reduction explores the question of what information is "important." The concept is very old; for example, we learned early on that in the analysis of a chemical reaction, the important variables often include temperature and concentrations, which are spatial averages of molecular speeds and positions. The reaction may involve collisions between individual molecules, but we only need to know the value of these average properties to predict how they will change. However, as we attempt to model new or more complex systems, we desire more systematic ways to extract these important variables from the available data. The most ubiquitous method is principal components analysis (PCA), which can be traced back to as early as 1901 [27], but was developed across many fields under many names over much of the 20^{th} century [16, 33]. PCA can be thought of as an extension of the line-of-best-fit method to higher dimensions: if the data is viewed as a cloud of points in some high dimensional space, PCA determines which directions characterize the highest variability in the data. Projecting the data onto only these directions results in a reduced representation of the data that still captures most of the variability. One drawback to PCA is that it only captures linear relationships (consider the case of a line-of-best-fit vs. a nonlinear curve fit). The field of nonlinear dimensionality reduction has grown rapidly in recent years with the introduction of methods like kernel PCA [30], local linear embedding [29], Isomap [36], Laplacian eigenmaps [1], and diffusion maps [4, 5, 3]. By capturing nonlinear structure, these methods can reduce data further than linear methods.

Dimensionality reduction methods have the potential to greatly improve clarity and speed of models by reducing the model to lower-dimensional representations that are easier to visualize and faster to compute. We are interested in using diffusion maps to improve model performance in molecular simulation and other relevant chemical engineering systems. Diffusion maps has recently shown promise in data analysis for models in transport [35], chemical kinetics [2], and molecular dynamics [10, 11, 26, 22]. We intend to expand on this work in the context of the "equation-free" (EF) framework, which uses short simulations of detailed models as the basis for projection in a coarse model (whose variables can be determined by diffusion maps). This idea of data analysis for the purpose of computation includes the idea of aiding in the development of a model or the characterization of dynamical systems by biasing where to next initialize simulations to gain "important" information. This in particular has applications to molecular simulation, where the characterization of rare

events can be extremely expensive; methods to speed up these computations often require a good coarse variable to bias the simulation.

This document will first explain the EF framework and then discuss diffusion maps and some of the challenges in its implementation. We will describe current work being done on a data-centric application in developmental dynamics, and then we will conclude with systems of interest in future work.

2 Equation-Free Modeling

EF modeling refers to the framework of using exisiting detailed microscopic simulators in a black-box manner to enable solutions to macroscopic tasks that are intractable for the microscopic simulator in its original formulation [31, 20, 21]. Consider such a simulator (e.g., molecular dynamics or computational fluid dynamics code) which performs detailed simulation (i.e, time-stepping on the microscopic level) given certain initial conditions and parameters. EF modeling requires the existence of an appropriate coarse (low-dimensional) description, where appropriate in this context means that it should capture the "important" information. We want the detailed simulation to track along some lower dimensional manifold (which can be parameterized by our coarse variables) embedded in the detailed, high-dimensional space. Simulations that are initialized away from this manifold should very quickly move to the manifold in the spirit of chemical kinetics systems where some species can be assumed to follow the quasi-steady-state approximation, which constitutes an effective dimensionality reduction.

EF modeling also requires that we be able to move between coarse (low-dimensional) and fine (high-dimensional) descriptions through some reasonably tractable operators. A restriction operator converts a fine description to a coarse description; for example, we might restrict the speeds of each molecule in MD simulation to the coarse variable temperature. Lifting operators, which do the reverse, often have the added complication that there are many fine descriptions that correspond to the same coarse description, so some care needs to be taken in systematically choosing a fine description. In many cases, however, even if the fine simulation is initialized poorly, it will quickly move back to the appropriate manifold. This quick relaxation is typically referred to as healing.

Given the detailed simulator, the appropriate coarse description, and the two operators, the most straightforward application of the EF framework is coarse projective integration (CPI), where we step the coarse variables forward in time based on properly initialized fine simulations. For each coarse integration step, the procedure is as follows (see Figure 1 for a schematic):

1. Use the lifting operator to determine initial conditions for the fine simulation based on the current values of the coarse variables.

2. Run the fine simulation for a short healing period to bring it back to the appropriate manifold.

3. Continue to run the fine simulation until sufficient information is captured about about the progression of the coarse variables.

4. Project the coarse variables forward in time using forward Euler integration or any other integration scheme.

If the simulation data actually lie on or near a manifold characterized by the coarse variables, coarse projective integration offers vast speed-ups compared to straight integration of the fine simulation. This framework of using fine simulations to determine the coarse behavior also has applications in stability/bifurcation analysis [37] [14] and the exploration of potential surfaces [13]. We are interested in applying this approach to systems where the coarse variables are not known *a priori* but must be determined through dimensionality reduction.

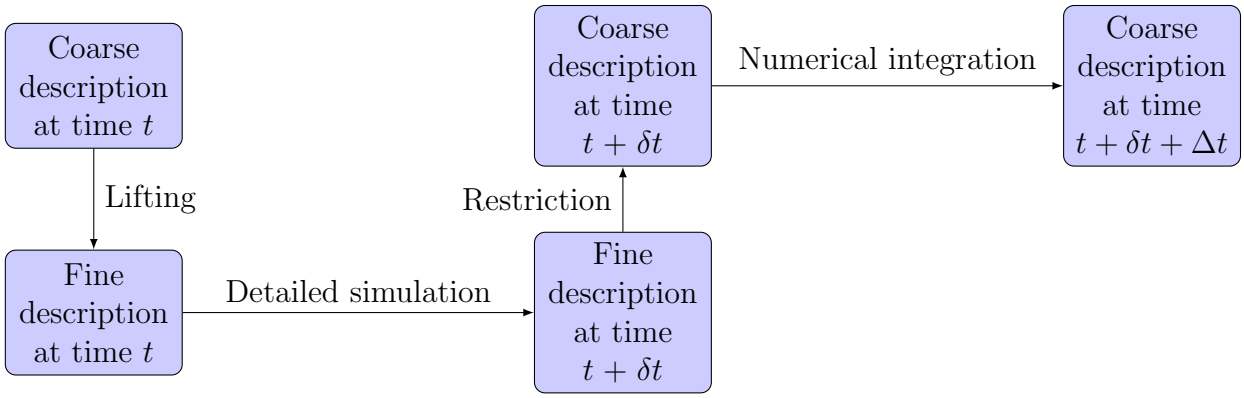

Figure 1: Schematic for a coarse projective integration step. We assume Δt can be chosen much larger than δt for significant speedup.

3 Diffusion Maps

3.1 Overview

Diffusion maps is a method of analyzing the geometry of data and discovering lower-dimensional manifolds that the data approximates (for an example, see Figure 2). The algorithm is designed to approximate the continuous Laplace-Beltrami operator (which has been shown to provide good parametrizations of nonlinear manifolds [19]) to discrete data. Suppose you have m data points in n-dimensional space represented by $y_1,...,y_m$. The algorithm first

constructs a weight matrix such that

$$W_{ij} = exp(-\frac{||y_i - y_j||^2}{\sigma^2})$$

where $|| \bullet ||$ is an appropriate norm or distance metric between two data points and σ is a characteristic distance such that points are considered close. Diffusion maps treats distances smaller than σ as important but treats distances much longer than σ as meaningless. The matrix is then made row-stochastic by dividing each row by its sum so that the rows sum to 1. This gives W the interpretation of a Markov matrix such that the elements represent transition probabilities from one data point to another, thereby modeling random walk diffusion among the data points. Variations to the algorithm exist which can, among other things, account for variations in sampling density.

The eigendecomposition of W yields eigenvalues λ_0,...,λ_{m-1} and eigenvectors ϕ_0,...,ϕ_{m-1}. Due to row-stochasticity, the first eigenvector ϕ_0 is a trivial constant vector with $\lambda_0 = 1$. The other eigenvectors provide a new coordinate system such that the k^{th} component of y_i is given by the i^{th} component of ϕ_k, scaled by λ_k. It has been shown that in this new coordinate system, Euclidean distance between two points approximates a concept commonly referred to as a diffusion distance. This diffusion distance represents moving from one point to another by random walk diffusion, where you can only move to nearby points based on the probabilities in the weight matrix. Since diffusion distance is based on moving from point to point, it approximates the distance along the manifold that the points lie on. If there is a large spectral gap, meaning that some eigenvalues are significantly larger than others, then the diffusion distance can be accurately approximated using only coordinates with the largest eigenvalues. The number of these large eigenvalues gives information about the true dimensionality of the manifold. One complication is that since the true eigenfunctions of the Laplace-Beltrami operator (the diffusion operator which we have approximated) are periodic, not all coordinates will parameterize new directions on the manifold, but will instead be higher harmonics. Methods have recently been proposed to automatically identify and remove these coordinates, which allows for maximum dimensionality reduction [6].

3.2 Choosing a distance metric and kernel scale

Since the diffusion maps algorithm is based entirely on the distances between points, choosing an approriate distance metric is critical. Assuming that each data point is represented by a vector such that each element represents one variable that describes the state (e.g., concentration of one chemical species, signal intensity at some point in space, etc.), then the obvious solution is simply the Euclidean distance between the two points described by the two data vectors. Unfortunately, in many instances, straightforward application of the Euclidean distance is not an informative measure of similarity. Here is a listing of some useful modifications and alternate approaches:

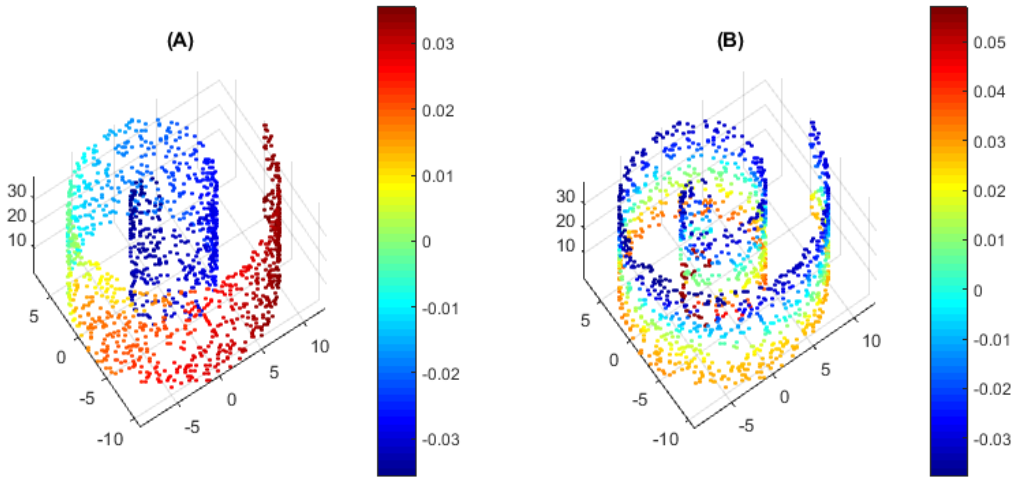

Figure 2: The canonical manifold learning example: the Swiss roll. 1500 points randomly sampled from a Swiss roll, colored by the (A) first and (B) second diffusion maps coordinates. The first coordinate paramterizes arc length along the swiss roll, while the second parameterizes the height. Although the original data is in three dimensions, it can be more simply described by just those two.

1. **Data preprocessing.** This is a general term for adjusting the data, either manually or systematically, so that the Euclidean distance becomes more informative. This can include actions like standardization of variables which span disparate scales, blurring of spatially organized data to remove uninformative small-scale structure, and contrast boosting or other image enhancement.

2. **Feature extraction.** In certain cases, *a priori* knowledge about the system can justify the selection of informative features. These functions of the original data can include ratios of quantities, times to achieve certain benchmarks, or quantities of distributions. For example, in a particular molecular simulation, it may already have been shown that important features include average kinetic energy per particle or the peaks in the radial distribution function. This is essentially an initial step of dimensionality reduction before allowing the formal algorithm to finish the job. Unfortunately, this *a priori* knowledge is often unavailable for new systems.

3. **Mahalanobis distance** [6, 8, 25]**.** This metric is primarily useful in stochastic systems where movement along the slow manifold is obscured by independent white noise. Based on the covariance of the data, the Mahalanobis distance ignores directions which do not indicate meaningful dissimilarity. Given column vector observations y_i and y_j

from a distribution with covariance C, the Mahalanobis distance is given by

$$||y_i - y_j||_M^2 = (y_i - y_j)^T C^{-1} (y_i - y_j)$$

The covariance matrix typically must be estimated from the data, and often the estimation is not full-rank so a pseudoinverse is required. For an example of where this is useful, see Figure 3.

4. **Earth mover's distance** [23]**.** Initial feature extraction sometimes motivates the use of distributions on each data point rather than the raw data (e.g., when each data point describes a collection of interchangeable particles). Also known as the 1^{st} Wasserstein metric, the earth mover's distance quantifies how much work is required to change one histogram into another; in the limit of infinite data and discretization, EMD compares probability distributions.

5. **Graph metrics.** A graph is a representation of data that defines nodes (which can be people, websites, locations, etc.) and edges (connections) between nodes. These edges can be either weighted or unweighted, and either directed or undirected. Methods for graph similarity include edit distance, maximal common subgraph, and graph kernels. These methods are often computationally intractable, but methods exist for their approximation.

Once the distance metric is chosen, the kernal scale σ must also be chosen. This sometimes requires trial and error, but a good initial guess is to use some fraction (e.g., $\frac{1}{2}$) of the median pairwise distance between data points. An alternate approach is to use the maximum distance to some number of nearest neighbors, averaged over each data point.

3.3 Synchronization of data: factoring out symmetry

Another issue that can occur in the application of diffusion maps is when data points have some degree of freedom among some symmetry group that obscures the dynamics. We would like to factor out the effects of symmetry and focus on the dynamics. One example is dynamics under periodic boundary conditions where we often want to factor out the translation. Other examples include 2D images that can be arbitrarily rotated, or molecular simulations of a complex molecule that can be arbitrarily shifted and rotated in 3D space. A common way to solve this problem is to align each image based on some template, but when the dynamics are complicated, it is not always obvious how to choose a template without *a priori* knowledge, and pairwise alignment between a data point and a template can be noisy. The eigenvector alignment method proposed by Singer [32] solves this problem by considering each data point as the template for each of the other points and finding a globally optimal alignment for each image.

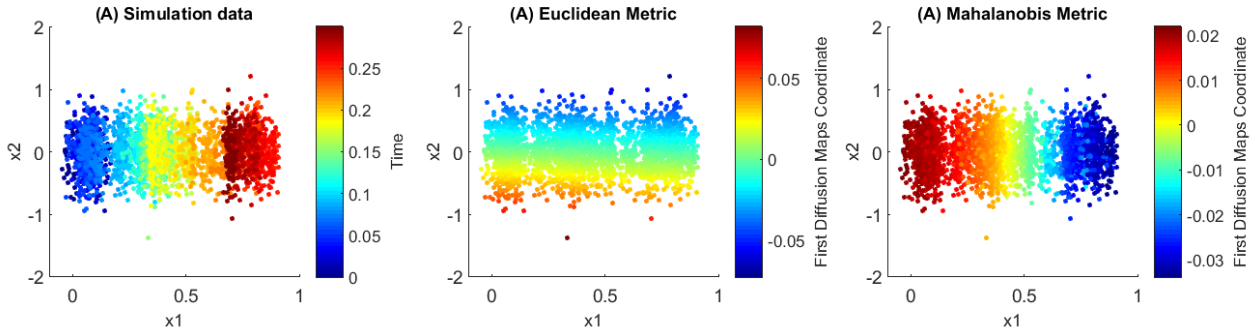

Figure 3: Benefits of the Mahalanobis distance. (A) Results of a stochastic simulation colored by time. The slow variable (x_1) gradually drifts from left to right over the course of the simulation. The dynamics are obscured by a fast variable (x_2) which has mean 0 but large Brownian motion. (A) Data points are colored by the first diffusion maps coordinate using the Euclidean distance. We recover the fast variable, which gives no information about the state of the system. (B) Data points are colored by the first diffusion maps coordinate using the Mahalanobis distance. We recover the slow variable, which is much more indicative of the state of the system. More analysis can be found in [6].

The algorithm requires that the transformation between members of the group be represented by some operator g_{ij} such that the operator satisfies the triplet consistency relation:

$$g_{ik} = g_{ij}g_{jk}$$

In the case of 1D periodic boundary conditions, we can consider the dynamics to be taking place on a ring, and then the operator g_{ij} becomes $exp(-\theta_{ij})$, where θ_{ij} is the shift that takes data point j to data point i. For general rotations, g_{ij} becomes the rotation matrix R_{ij}, which can be 2 by 2 for rotations in the plane or 3 by 3 for rotations in $\mathbb{R}^3$ [34]. Once this operator is defined, we construct the matrix G such that each the ij^{th} element (or ij^{th} block) of G is the operator g_{ij}. The top eigenvector (or eigenvectors in the block matrix case) gives the operators g_i which transformed the optimal representation into data point i. This representation has degrees of freedom within the group (in the case of rotation within the plane, the solutions could be arbitrarily rotated within the plane), but it is optimized based on all pairwise comparisons.

An extension called vector diffusion maps (VDM) exists which simultaneously aligns the images while performing diffusion maps. VDM is useful in cases where the dynamics are dramatic, and pairwise alignments between dynamically different data points become meaningless. VDM ignores these alignments and simultaneously provides global alignment information and dimensionality reduction [34, 8].

3.4 Lifting and restriction operators

Various choices exist for lifting and restriction operators, which are essentially interpolation/curve fitting schemes. Several choices are summarized here; more detailed explanations and analysis can be found in [2], and more specific sources are also given below. The problem statement is as follows: given a set of detailed descriptions y_1,...,y_m, along with a set of corresponding coarse descriptions ϕ_1,...,ϕ_m from diffusion maps, how do we assign a new coarse description ϕ to a new detailed state y (restriction) and vice versa (lifting).

1. **Nearest Neighbor.** The simplest approach is to take some an average of some number of nearest neighbors, possibly with some weighting based on distance to each neighbor. A slightly more sophisticated approach is to fit functional (often linear) relationships between the coarse and fine variables based on just the nearest neighbors.

2. **Nyström Extension** [12]**.** This is one of the most common methods for finding diffusion maps coordinates for a new detailed state. The new coordinate is essentially calculated as a weighted sum of all of the previous coordinates with weights based on the diffusion kernel. The equation for the α^{th} component of the new coarse description is given by

$$\phi_\alpha = \frac{1}{\lambda_\alpha} * \frac{\sum_{i=1}^{m} \phi_{i,\alpha} * exp(-\frac{||y-y_i||^2}{\sigma^2})}{\sum_{i=1}^{m} exp(-\frac{||y-y_i||^2}{\sigma^2})}$$

   The presence of $\frac{1}{\lambda_\alpha}$ ensures that applying the procedure to the existing points gives the same coordinates as the original diffusion maps algorithm; this can be verified by considering the eigenvalue problem in section 3.1.

   In its typical formulation, the Nyström extension is only a restriction operator, but it is possible to consider a lifting self-consistency problem of finding a new data point that would give a specific coarse description from Nyström. While this has not been well-studied, it would likely involve opimization of an initial guess based on another method.

3. **Radial Basis Functions** [28]**.** This method can be implemented based on any function that depends only on the distance between two input points. Both lifting and restriction operators can be expressed on a sum over nn nearest neighbors:

$$\phi_\alpha = \sum_{i=1}^{nn} \beta_{i,\alpha} * f(||y - y_i||)$$

   The coefficients β can be determined using linear algebra based on the nn fitting points. The Nyström extension can be thought of as a modified specific case of RBF where the radial function f is the diffusion kernel and each β involves λ_α as well as the normalization constant (which depends on y, unlike in RBF). Other common basis functions include simple powers of the distance.

4. **Kriging.** Also known as Gaussian Process Regression, this method views the functional relationship as a realization of a stochastic Gaussian process based on the assumption that points that are nearby in space are statistically correlated. The first step is to fit (based on the data) a model for the semivariogram

$$\gamma(y_i, y_j) = var(\phi(y_i) - \phi(y_j))$$

where var(•) is the variance. The semivariogram is a measure of how different the coarse descriptions at y_i and y_j are likely to be; it often taken to be simply a function of the distance, although modifications exist to account for variation due to location and orientation in fine space. Depending on the assumptions, the Kriging equations can be slightly different: good discussions can be found in [28] and [18].

5. **Laplacian Pyramids** [9]**.** This is a multiscale method which finds ϕ_α as a sum of fits on smaller and smaller scales. Assuming a given kernel k (typically the normalized diffusion kernel from Nyström using smaller and smaller σ), the first fit is found by

$$s_\alpha^{(0)} = \sum_{i=1}^{nn} k^{(0)}(y_i, y) * \phi_{i,\alpha}$$

Subsequent fits are found by reducing the kernel scale and fitting to the residual:

$$s_\alpha^{(l)} = \sum_{i=1}^{nn} \left(k^{(l)}(y_i, y) * (\phi_{i,\alpha} - \sum_{j-0}^{l-1} s_\alpha^{(j)}) \right)$$

After evaluating to an appropriate scale l_{max}, the final solution is given by

$$\phi_\alpha = \sum_{l=0}^{l_{max}} s_\alpha^{(l)}$$

Alternative multiscale methods exist, such as geometric harmonics.

3.5 Extending outside of the support

There is a nontrivial issue that arises when attempting to do calculations with diffusion maps coordinates that extend outside the support of the training data. The diffusion maps eigenproblem is an approximation of the Laplace-Beltrami operator on an underlying manifold. In the typical framework, diffusion only takes place between the data points, which implies no flux at the boundary. In the continuous case, the eigenfunctions involve cosines, so the slope at the boundary is 0. This means that attempting to model the behavior of diffusion maps coordinates outside the boundary is ill-posed; the diffusion maps coordinate is not changing at the boundary, and finding a fine state with a diffusion maps coordinate

greater than the maximum is not possible for linear fit. Even in the discrete approximation where the slope doesn't go fully to 0, extension past the boundary is extremely error prone.

It is desirable to eliminate the zero slope at the boundary by modifying the eigenproblem to approximate boundary conditions with a specified flux. While some work has been done in this area, it remains more of an art than a science. It is necessary to first identify boundary points in each coarse direction. Row-stochasticity is eliminated for these points to eliminate the no-flux boundary condition. This allows much better modeling and extrapolation near the boundary.

4 Application to Data Fusion: *Drosophila* embryos

4.1 Data description and motivating problem

One application of diffusion maps is in the study of developmental dynamics. This application is not within the EF framework and is more focused on the data itself rather than using the data for calculations, but it does showcase many of the relevant tools in data analysis that can be used to support EF modeling. Here, the specific case is the development of *Drosophila* embryos, where researchers are interested in tracking the embryo structure (as represented by the locations of cell nuclei) as well as the distribution of various relevent proteins [24] [7]. There are two relevant types of data sets:

1. A live movie set consists of images every 30 seconds of the same embryo over some range of its development time. These images have nuclei labeled by Histone-RFP, and also have associated time stamps. See Figure 4.

2. Fixed snapshot data sets consist of one image from each of a set of embryo. These embryo have three channels; they are stained different colors for the nuclei as well as two proteins, Twist and dpERK. These images do not have time stamps, and they may also be arbitrarily rotated in the plane. See Figure 5.

The motivating problem is to construct an average developmental trajectory that contains information about all three channels as well as approximate timings. Prior work in the group has shown that the vector diffusion maps algorithm can very accurately reconstruct the ordering of the live movies even when they are scrambled and arbitrarily rotated. It has also been shown that VDM can correctly rotate a data set of fixed snapshots, and the ordering is reasonably well consistent with manual ordering by an expert. The next step is to fuse these data sets together to produce a trajectory with color and time.

4.2 Challenges in data preprocessing

Each live image consists of 100 by 100 pixels; fixed snapshots are more but are subsampled. The live movies have just one channel value at each pixel, while the fixed snapshots have

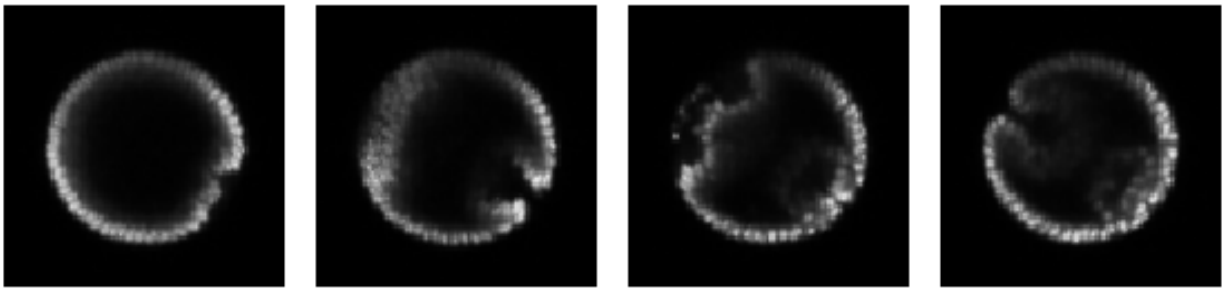

Figure 4: Sample snapshots from a single live movie taken at 5, 10, 15, and 20 minutes. Note that the embryo stays in the same orientation.

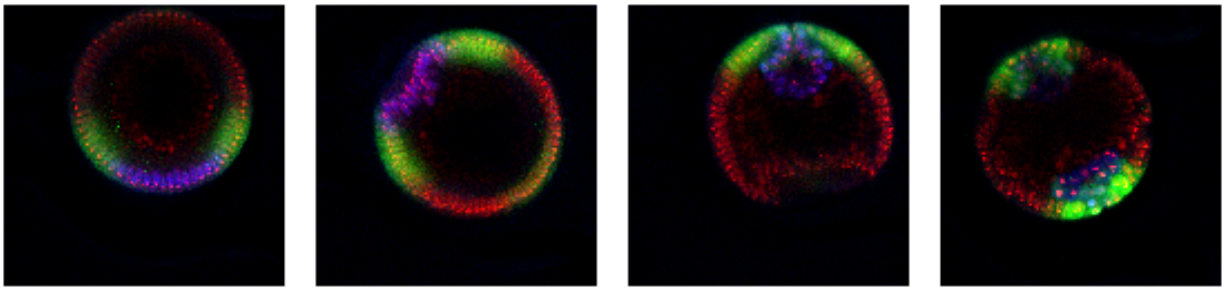

Figure 5: Sample fixed snapshots. Red is nuclei, while blue and green represent particular proteins. The times and true orientation are not known *a priori*, and the images may be translated or dilated.

three. Each image can thus be viewed as a vector of length $100,000$ (or $300,000$ for fixed snapshots), and a simple first choice of metric is the Euclidean distance between two such vectors. Unfortunately, some significant work needs to be done to make that Euclidean distance informative. The imaging process can produce various image artifacts that are not relevent to the developmental dynamics, especially for the fixed snapshots which are all different embryo. Images need to be centered in the field of view and resized to occupy a consistent portion of the frame. Contrast limited adaptive histogram equalization (CLAHE) is used to normalize the images. We recently found that boosting the contrast significantly improves the performance, probably because whether or not a pixel is occupied is more important than the specific brightness level. Finally, we blur the images using a Gaussian filter to remove small scale structure that distorts the comparison. Variability between embryo means that features will not be in the exact same place; blurring allows such features to be more properly compared.

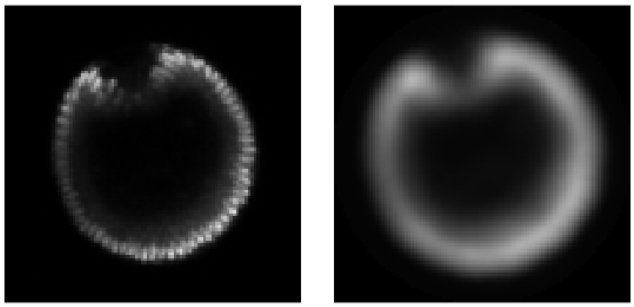

Figure 6: Example raw image and preprocessed image.

4.3 Synchronization of rotations and movie times

While the embryos in the live movies stay at the same orientation throughout development, the embryos in the fixed snapshots can be in any arbitrary orientation within the plane. While manual registration is possible, it is extremely tedious. The eigenvector alignment method is capable of registering the fixed snapshots with respect to each other in just seconds of computational time. As previously discussed, the final orientation is arbitrary, so there is a final step of rotating the snapshots to align with the live movies in order to do any comparison. The eigenvector alignment method also provides a framework for aligning the different live movies and the snapshots with respect to each other, but with only 7 live movies and 1 data set of snapshots, the gains over manual registration are minimal.

Another synchronization of interest is that of movie times. When diffusion maps is run on multiple live movies simultaneously, we find that within each live movie there is a good monotonic relationship between time and the first diffusion maps coordinate. However, these different functional relationships are offset from each other. We believe that while some of this error is due to noise and embryo variability, issues with experimental design could be causing some of the discrepancy. For example, lack of precision in the fertilization time could mean that the times for one embryo are shifted in relation to another. Differences in ambient temperature or other environmental factors could cause one embryo to develop at a faster rate. We are interesting in finding factors by which to scale and shift the movie times so that they optimally align with each other. If we consider operations of the form

$$t_j = a_{ij} * t_i + b_{ij}$$

(i.e., how to scale and shift the times in movie i to align with movie j), then the eigenvector alignment method can be used by considering the matrix operator

$$g_{ij} = \begin{bmatrix} 1 & 0 \\ a_{ij} & b_{ij} \end{bmatrix}$$

The pairwise a_{ij} and b_{ij} can by quickly found by minimizing the least squares error between the two relationships for time as a function of the first diffusion maps coordinate. We

then form the block matrix G, and the top two eigenvectors can be used to form a block eigenvector which contains the scales and shifts from the globally optimal solution. Again, there are degrees of freedom in the scale and in the shift, so we apply a final scale and shift such that the mean shift that we apply to the data is 0 and the mean log scale we apply is also 0. This is in an effort to modify the data as little as possible, and also to make the algorithm deterministic.

4.4 Data fusion to construct a representative trajectory

Our data fusion methodology is based on the assumption that the underlying process of embryonic development is primarily one-dimensional and that the primary dimension is monotonic with time. Different methods of observation (i.e., different imaging techniques) embed this 1D process in a higher-dimensional space. Imaging noise and inter-embryo variability represent high-dimensional noise around this one-dimensional manifold.

The initial attempt at data fusion used the additional assumption that the two data sets (live movies and fixed snapshots) had common information: the channel representing nuclei locations. At this point several algorithms are possible; we considered the following:

1. Run diffusion maps on common information from the (preprocessed) live movies and fixed snapshots as one data set.

2. Based on the data, develop functional relationships between color values at each pixel and the diffusion maps coordinate.

3. "Color" the live movies based on these functional relationships.

4. Apply some smoothing to the colored live movies to create a smooth, representative trajectory.

The results from this algorithm unfortunately failed multiple sanity checks. The functional relationships showed rough trends but were unreasonably noisy. The ordering of the fixed snapshots in diffusion maps space was poorly correlated with the ordering obtained by an expert. The fixed snapshots spanned a significantly wider range in diffusion maps space than the live movies did.

While the general noisiness of the system contributes to these problem, another significant problem is comes from difference in imaging techniques. Visual inspection shows clear differences between the snapshots and the live movies, most notably in signal strength in the yolk of the embryo. This is believed to be related to the fact that different nuclei staining methods were used. Because of these issues, we violate the assumption that the two data sets have common information; instead they represent different observation functions of the common information.

One approach to this problem is more manual preprocessing of the fixed snapshots. If they can be adjusted so that the (hidden) observation functions are the same, the algorithm should have more success. We have experimented with removing the signal from the yolk, and the results show significant improvement but are still unsatisfactory. We are still investigating alternative preprocessing methods.

Another approach involves transforming the diffusion maps coordinates of the snapshots to better align with the coordinates of the movies. The challenge is to do so in a robust way that maintains the intra-data-set information while aligning the seperate data sets. We may include more diffusion maps coordinates (i.e., a higher dimensional representation), and consider transoformations that translate and twist but do not stretch. Both approaches are in the stage of manual experimentation; we hope to automate as much as possible when a viable approach is found.

5 Systems of Interest for Future Work

This section is focused on systems where we hope to use diffusion maps to enable calculations in the EF framework.

5.1 Molecular simulation

The utility of molecular simulation is often constrained by computational limitations. Great strides have been made to circumvent these restrictions through techniques like parallel tempering and umbrella sampling, but researchers are often still limited to very short times and oversimplified molecular models. We hope to improve the speed of these simulations with more robust and systematic model reduction.

An area of molecular simulation that has shown promise for our techniques is protein dynamics and folding. Alanine dipeptide (Ala2) is a common reference molecule for testing of new techniques. It has been well studied experimentally and computationally [38], and various studies have shown that its dynamics can be well described using a few coarse variables. In 2003, Hummer et. al. implemented several methods of equation-free analysis based on one of the dihedral angles as the coarse variable [17]; in 2009, Frewen et. al. used similar techniques to explore the potential energy landscape using two of the dihedral angles [13]. Other work has shown that diffusion maps can characterize the dynamics and find good coarse variables. Nedialkov et. al. recently used diffusion maps to parametrize the more complicated alanine pentapeptide, and used the results in clustering algorithm to develop a Markov model [26]. Ferguson et. al. showed that application of diffusion maps can recover the dihedral angles as coarse variables, and proposed a method to iteratively incorporate diffusion maps variables into successive rounds of umbrella sampling [11].

We would like to combine the diffusion maps approach with the EF framework and show that diffusion maps can provide good coarse variables for EF methods to use. One issue

noted in the Ferguson work is that their is no systematic method for determining physical variables for umbrella samplling that the diffusion maps variables parameterize. However, in the EF framework, the functional relationship between diffusion maps coordinates and physical parameters only needs to be known on the scale of the coarse time step, not globally. This should be feasible using the lifting and restriction operators described in section 3.4.

5.2 Reaction networks

Diffusion maps has been shown to parameterize reaction networks represented by systems of ODEs [2]. This is one of the few areas where diffusion maps has been used to improve calculations rather than clarity; Chivazzo et. al. used their work to develop a reduced model using lifting and restriction operators. However, they still used the framework of data collection followed by model formulation, and they note that improvements could be made by using the diffusion maps coordinates to extend the manifold. The main computational cost of constructing the model involves sufficiently sampling the manifold; this could potentially be reduced by using the diffusion maps coordinates to predict which new evaluations will give more information about parts of the manifold that haven't yet been explored.

Diffusion maps can also be applied to reaction networks described by stochastic simulation rather than continuous ODEs [9].The Gillespie stochastic simulation algorithm (SSA) is an efficient method for simulating reactions in systems with relatively small numbers of molecules, where the stochastic nature of reaction kinetics is not averaged out [15]. Based on the number of molecules in the system and the rate constants of possible reactions, SSA calculates probability distributions for each reaction occuring in a given time window. Based on these distributions, a reaction and time to reaction are selected randomly; the SSA procedes forward by repeatedly selecting the next reaction to take place. The stochastic nature of the simulation suggests the use of the Mahalonobis distance in diffusion maps to factor out the noise and discover the slow manifold that systems drifts on.

6 Conclusions

The equation-free framework has been demonstrated to improve computational tractability in detailed systems that can be effectively parameterized by a low-dimensional, coarse description. This methodology depends primarily on discovering the coarse variables as well as efficient lifting and restriction operators to move between coarse and fine variables. A major stumbling block is the ability to determine these coarse variables systematically without *a priori* information about the dynamics.

Diffusion maps is a manifold learning method that has been shown to effectively discover low dimensional descriptions of complex data in systems ranging from developmental biology to molecular dynamics. To this point, it has been used primarily to improve clarity of the dynamics. Only very recently has work been done using diffusion maps to improve the

efficency of calculations, but methods to this point have relied either on fully collecting the data set before hand, or on manual correlation of diffusion maps coordinates with globally applicable physical variables.

The combinations of these approaches has applications in all fields of computational modeling. We hope to demonstrate the viability of diffusion maps to enable use of equation-free modeling in molecular dynamics and reaction kinetics.

References

[1] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, 2003.

[2] E. Chiavazzo, C. Gear, C. Dsilva, N. Rabin, and I. Kevrekidis. Reduced Models in Chemical Kinetics via Nonlinear Data-Mining. *Processes*, 2(1):112–140, Jan. 2014.

[3] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[4] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–31, May 2005.

[5] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7432–7, May 2005.

[6] C. J. Dsilva. *Manifold Learning for Dynamical Systems*. PhD thesis, Princeton University, 2015.

[7] C. J. Dsilva, B. Lim, H. Lu, A. Singer, I. G. Kevrekidis, and S. Y. Shvartsman. Temporal ordering and registration of images in studies of developmental dynamics. *Development*, pages 1717–1724, 2015.

[8] C. J. Dsilva, R. Talmon, C. W. Gear, R. R. Coifman, and I. G. Kevrekidis. Data-driven Reduction for Multiscale Stochastic Dynamical Systems. xx:1–19.

[9] C. J. Dsilva, R. Talmon, N. Rabin, R. R. Coifman, and I. G. Kevrekidis. Nonlinear intrinsic variables and state reconstruction in multiscale simulations. *The Journal of chemical physics*, 139(18):184109, 2013.

[10] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 107(31):13597–602, 2010.

[11] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis. Integrating diffusion maps with umbrella sampling: Application to alanine dipeptide. *The Journal of Chemical Physics*, 134(13):135103, 2011.

[12] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nystrom method. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

[13] T. A. Frewen, G. Hummer, and I. G. Kevrekidis. Exploration of effective potential landscapes using coarse reverse integration. *Journal of Chemical Physics*, 131(13), 2009.

[14] C. W. Gear, I. G. Kevrekidis, and C. Theodoropoulos. 'Coarse' integration/bifurcation analysis via microscopic simulators: micro-Galerkin methods. *Computers & Chemical Engineering*, 26(7–8):941–963, 2002.

[15] D. T. Gillespie and D. T. Gillesple. Exact Stochastic Simulation of couple chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[16] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[17] G. Hummer and I. G. Kevrekidis. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *Journal of Chemical Physics*, 118(23):10762–10773, 2003.

[18] E. H. Isaaks and R. M. Srivastava. *Applied Geostatistics*. Oxford University Press, New York, 1989.

[19] P. W. Jones, M. Maggioni, and R. Schul. Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):1803–8, 2008.

[20] I. Kevrekidis and C. Gear. Equation-free, coarse-grained multiscale computation: Enabling mocroscopic simulators to perform system-level analysis. *Communications in Mathematical Sciences*, 1(4):715–762, 2003.

[21] I. G. Kevrekidis, C. W. Gear, and G. Hummer. Equation-free: The computer-aided analysis of complex multiscale systems. *AIChE Journal*, 50(7):1346–1355, July 2004.

[22] S. B. Kim, C. J. Dsilva, I. G. Kevrekidis, and P. G. Debenedetti. Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein. *The Journal of Chemical Physics*, 142(8):085101, 2015.

[23] E. Levina and P. Bickel. The earth mover's distance is the Mallows distance: some insights from statistics. *Eighth IEEE International Conference on Computer Vision*, pages 251–256, 2001.

[24] B. Lim, C. Dsilva, T. Levario, H. Lu, T. Schüpbach, I. Kevrekidis, and S. Shvartsman. Dynamics of Inductive ERK Signaling in the Drosophila Embryo. *Current Biology*, pages 1–7, 2015.

[25] P. C. MAHALANOBIS. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[26] L. V. Nedialkova, M. A. Amat, I. G. Kevrekidis, and G. Hummer. Diffusion maps, clustering and fuzzy Markov modeling in peptide folding transitions. *The Journal of chemical physics*, 141(11):114102, Sept. 2014.

[27] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(1):559–572, 1901.

[28] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes*. Cambridge University Press, New York, third edition, 2007.

[29] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[30] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998.

[31] C. I. Siettos, C. C. Pantelides, and I. G. Kevrekidis. Enabling Dynamic Process Simulators to Perform Alternative Tasks: A Time-Stepper-Based Toolkit for Computer-Aided Analysis. *Industrial & Engineering Chemistry Research*, 42(26):6795–6801, Dec. 2003.

[32] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011.

[33] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America. A, Optics and image science*, 4(3):519–524, 1987.

[34] B. Sonday, A. Singer, and I. G. Kevrekidis. Noisy dynamic simulations in the presence of symmetry: Data alignment and model reduction. *Computers & Mathematics with Applications*, 65(10):1535–1557, May 2013.

[35] B. E. Sonday, M. Haataja, and I. G. Kevrekidis. Coarse-graining the dynamics of a driven interface in the presence of mobile impurities: Effective description via diffusion maps. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(3):1–11, 2009.

[36] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500):2319–2323, 2000.

[37] C. Theodoropoulos, Y.-H. Qian, and I. G. Kevrekidis. "Coarse" stability and bifurcation analysis using time-steppers: A reaction-diffusion example. *Proceedings of the National Academy of Sciences*, 97(18):9840–9843, 2000.

[38] H. Wang, C. Schütte, G. Ciccotti, and L. Delle Site. Exploring the conformational dynamics of alanine dipeptide in solution subjected to an external electric field: A nonequilibrium molecular dynamics simulation. *Journal of Chemical Theory and Computation*, 10(4):1376–1386, 2014.