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 inherently messy; 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. The reaction may involve collisions between individual atoms, but we only need to know certain things about the average behavior to predict the results that we are interested in. However, as we encounter new systems (which are often more complex), we desire more systematic ways to extract these important variables. The most ubiquitous method is principal components analysis (PCA), which can be traced back to as early as 1901 [13], but was developed across many fields under many names over much of the 20^{th} century [8]. 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 [15], local linear embedding [14], Isomap [18], and diffusion maps [3] [4] [2].

Dimensionality reduction methods have the potential to greatly improve models in terms of clarity and speed by reducing the model to the 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 [17], chemical kinetics [1], and molecular dynamics [5] [12] [11]. 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 that diffusion maps can aid in the development of a model or the characterization of dynamical systems by biasing where to next take data 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 [16] [9] [10]. Consider such a simulator (e.g., MD or CFD code) which performs detailed time-stepping on the microscopic level given certain initial conditions and parameters. EF modeling requires the existence of an appropriate coarse (lower 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 represented by our coarse description) 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 that follow the quasi-steady-state approximation.

EF modeling also requires that we be able to move between coarse (low dimensional) and fine (high dimensional) descriptions through some reasonable 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 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 integration step, the procedure is as follows:

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 [**?**] [7] and the exploration of potential surfaces [6]. 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.

3 Diffusion Maps

Diffusion maps is a method of analyzing the geometry of data and discovering lower-dimensional manifolds that the data approximates. The algorithm is designed to approximate the continuous Laplace-Beltrami operator (which has been shown to provide good parametrizations of nonlinear manifolds) 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 to 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. 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. In this new coordinate system, distance between two points is referred to as the diffusion distance. This diffusion distance represents moving from one point to another by 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 eigenvalues gives information about the true dimensionality of the manifold.

4 Application to Data Fusion: *Drosophila* embryos

5 Systems of Interest

6 Conclusions

References

[1] 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.

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

[3] 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.

[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: multiscale methods. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7432–7, May 2005.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.