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 [9], but was developed across many fields under many names over much of the 20th century [6]. 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 [11], local linear embedding [10], Isomap [13], 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 [12], chemical kinetics [1], and molecular dynamics [5] [8] [7]. 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

3 Diffusion Maps

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

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.