STATE REDUCTION IN MOLECULAR SIMULATIONS

Yuzhen Xue[1], Peter J. Ludovice[1], Martha A. Grover[1*], Lilia V. Nedialkova[2], Carmeline J. Dsilva[2] and Ioannis G. Kevrekidis[2*]
[1]School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA 30332
[2] Department of Chemical & Biological Engineering, Princeton University, Princeton, NJ 08544

Abstract

Model reduction is an important systems task with a long history in traditional chemical engineering modeling. We discuss its interplay with modern data-mining tools (such as Local Feature Analysis and Diffusion Maps) through illustrative examples, and comment on important open issues regarding applications to large systems arising in molecular/atomistic simulations.

Keywords

Model Reduction, Data Mining, Principal Component Analysis, Local Feature Analysis, Diffusion Maps.

Introduction

Model reduction is an important systems task in physics and engineering modeling, and it has a long history in chemical engineering marked by many successes over the years; the development of reduced chemical mechanisms for combustion and the lumping of petroleum fractions are two cases in point, and the use of these tools in biological reaction network reduction is a vital current research frontier. Mathematical tools ranging from the quasi-steady-state approximation and the partial equilibrium approximation to the center manifold theorem, Lyapunov-Schmidt reduction and Approximate Inertial Manifolds have been developed. Tools of a more computational nature, such as the Intrinsic Low-Dimensional Manifold approach or the Computational Singular Perturbation approach have also been finding extensive applications. We are interested in the reduction of large problems associated with molecular simulations; more specifically, we focus on data-mining tools that can be linked with model reduction techniques. We use extensions/modifications (both linear and nonlinear) of Principal Component Analysis to illustrate important open issues that we discuss.

Recent research by part of our team has studied the reduction of macromolecular simulations. The macromolecular example we will discuss here is a simple one, that aims to illustrate the features of the different data reduction methods employed. However, the motivation for reduction is derived from the complexities of molecular systems consisting of much larger molecules such as proteins, DNA and various synthetic polymers. At the atomic level the often important conformational modes for such molecules involve backbone torsion angles, and this will be illustrated in our dipeptide example. As these polymers become larger their simulation becomes rapidly prohibitive both because of their size and the longer simulations dictated by the increases in molecular relaxation times. Luckily, many large polymers have dominant dynamic modes that may be mapped to reasonable reduced coordinate systems. The methods described here – Local Feature Analysis (LFA), that allows one to target physically relevant modes by choosing "seed elements" systematically, and Diffusion Maps (DMAPs), that provide nonlinear parametrizations of these modes– attempt to construct such coordinate systems.

Examples of polymer systems and behaviors that may benefit from the methods described here include reptation for amorphous melts or glasses. Reptation is the slithering of a polymer chain along its own chain contour (Daoud and De Gennes, 1979). The primary coordinate for such a motion is along the contour of the polymer chain, similar to the artificial spiral data set we discuss here, for which Diffusion Maps are so effective. Many polymers have heterogeneous structure and behavior at spatial scales larger that the atomic scale; these include complex biological polymers such as proteins and RNA, for which flexible loops behave differently than more rigid helical motifs. This heterogeneity lends itself to systematic choices of seed elements for the LFA approach. We expect that other polymer systems, containing levels of structural order that are intermediate between an amorphous glass and a regular protein motif, may also benefit from such methods.

Technical Approach

Principal Component Analysis (PCA) has, for more than a century, held pride of place in the mining of data related to model reduction; its applications range from chemistry to turbulence and from sociology to biology. Starting with PCA we discuss and illustrate both linear (Local Feature Analysis)

[*]To whom all correspondence should be addressed:
martha.grover@chbe.gatech.edu, yannis@princeton.edu

and nonlinear (Diffusion Maps) extensions/modifications of it; this gives us the opportunity to discuss/highlight certain important current research issues.

Principal Component Analysis

Consider an ensemble $X \in \mathbb{R}^{N\times L}$ of L data points (vectors in $\mathbb{R}^N$ that have been centered, $L > N$). Let

$$R = \frac{1}{L-1} XX^T, \tag{1}$$

$R \in \mathbb{R}^{N\times N}$, denote the covariance matrix of this ensemble. It is well known the matrix R can be written as

$$R = \Psi_f \, \Lambda_f \Psi_f^T. \tag{2}$$

Λ_f is a diagonal matrix composed of R's eigenvalues in decreasing order: $\lambda_1 \geq \cdots \geq \lambda_r >> \lambda_{r+1} \geq \cdots \geq \lambda_N$, and Ψ_f is an orthogonal matrix composed of R's eigenvectors as columns. We define $\Psi \in \mathbb{R}^{N\times r}$ as the matrix containing the first r columns of Ψ_f, which correspond to the first r eigenvalues of R. PCA (Shlens, 2005) maps any vector $x \in \mathbb{R}^N$ down to $\tilde{x} \in \mathbb{R}^r$ through

$$\tilde{x} = \Psi^T x, \tag{3}$$

where $\tilde{x}$, the "output", contains the projection coefficients of x onto the first r principal components. In PCA, the projection coefficients are functions of all N variables, and therefore reflect "global" features. The inverse mapping (reconstruction) of PCA is defined by

$$x \approx \Psi\tilde{x}. \tag{4}$$

Local Feature Analysis

Local Feature Analysis is a linear dimensionality reduction method, similar to (and derived from) PCA (Penev and Atick, 1996). As in PCA, LFA considers the covariance matrix $R = \frac{1}{L-1}XX^T$. In PCA, the mapping from N down to r dimensions is accomplished using Ψ^T, as defined above. In LFA the matrix that maps the original N-dimensional data to a reduced r-dimensional space is based instead on the *topographic* kernel

$$K = \Psi\Lambda^{-\frac{1}{2}}\Psi^T. \tag{5}$$

Here $K \in \mathbb{R}^{N\times N}$, and the reduction is performed by the matrix $\tilde{K} \in \mathbb{R}^{r\times N}$, which consists of r judiciously selected linearly independent rows of K. Thus $\tilde{x} = \tilde{K}x$, where $x \in \mathbb{R}^N$ is a data point in the original state space and $\tilde{x} \in \mathbb{R}^r$ is its representation in the reduced subspace. The PCA and LFA methods each also give rise to an approximate inverse mapping, such that any point in the r-dimensional space can be "reconstructed" in the N-dimensional space. When the same value of r is used in PCA and LFA, the reconstruction accuracy obtained from both methods is identical (Penev and Atick, 1996).

In PCA, the rows of the mapping matrix are eigenvectors of R and are therefore orthogonal. In LFA, the rows of K are *not* orthogonal, but rather *contain topographic information*. In fact, any r independent rows of K can be used to represent the system, and thus can be selected for $\tilde{K}$. However, one can systematically select the r representative state variables, i.e. the "seeds". Many iterative methods (Penev and Atick, 1996; Zhang and Wriggers, 2008) have been proposed to automate the seed selection. While these proposed techniques use different criteria, they all involve significant computations and may not converge to a unique solution.

In our seed selection method all state variables are mapped, based on the data ensemble, onto a feature space spanned by the r principal eigenvectors. Note that in this feature space, all the eigenvectors are treated as being equally important (due to the rescaling in Eq. 5). A candidate pool of seed states is first constructed by selecting the states that significantly reflect the principal components, i.e. the states that have significant components in at least one of the principal directions. Among the candidate pool, the state that has the highest variance is selected as the first seed. That is, the state variable with index $\arg_i \max(\bar{\psi}_i \Lambda \bar{\psi}_i^T)$, where i is any index from the candidate pool and $\bar{\psi}_i$ is the i^{th} row in Ψ. The succeeding seed is selected from the remaining states in the pool so that the correlation between the current seed and the seeds selected so far is minimized. Here the correlation between state variables i and j can be shown to be indexed by the angle α between the normalized vectors of $\bar{\psi}_i$ and $\bar{\psi}_j$, where $0^\circ \leq \alpha \leq 90^\circ$. The smaller α is, the more correlated state variables i and j are. The termination condition is that either (a) there is no remaining state that is uncorrelated enough to the selected seeds or that (b) r "seeds" have been selected. The selected seeds are the least correlated in the PCA feature space and would span the LFA reduced dimensional state space. In fact, the number of seeds indicates the smallest number of original state variables that can be used to approximate the full system. If only $n < r$ seeds are selected in the end, the reduced dimensional state space spanned by these n seeds is different from the subspace spanned by the first n PCA eigenvectors. We remark that an alternative seed selection method can be found in our ongoing work (Xue et al., in preparation).

If reconstruction accuracy is the only consideration, there is no advantage of LFA compared to PCA. However, if the reduced-order coordinates are to be assigned physical meaning, then LFA can be advantageous. For a system characterized by strongly correlated topological groups, one expects LFA to pick up one representative state variable, i.e. seed, out of each group. In general, the consistency of the LFA basis has been reported to be higher than PCA, when applied to different windows of noisy data (Balsera et al., 1996; Zhang and Wriggers, 2008; Xue et al., 2010). Moreover -and this is a crucial point- associating coordinates in the reduced-dimensional space with a few specific variables in the original, N-dimensional space, may aid in physical interpretation of the reduction. In one of our examples, we will consider a molecular dynamics simulation, and choose seed atoms using LFA. This lays the foundation of an approach for automated coarse-graining of molecular simulations, in which each seed atom forms the locus for a "super-atom" in the coarse-grained simulation.

Diffusion Maps

In contrast to PCA and LFA, Diffusion Maps is a nonlinear dimensionality reduction technique (Coifman et al., 2005b,a). DMAPs take high-dimensional data points that lie on (or close to) a low-dimensional, nonlinear manifold and embed them in a low-dimensional linear space: the distance between data points *along the manifold* is related to the Euclidean distance in the new, embedding space, so that, in some sense, the intrinsic geometry of the data is retained. One constructs the $L \times L$ matrix W, with

$$w_{ij} = k(x_i, x_j) = exp\left(-\frac{d^2(x_i, x_j)}{\varepsilon}\right), \tag{6}$$

where k is a kernel function, d is a distance metric between the data points (in our examples we use the Euclidean distance), and ε, the kernel width, is a procedure parameter. This matrix W can be interpreted as the adjacency matrix for a weighted graph, where the nodes of the graph represent data points and two data points are connected by a high weight edge if they are close in the original space.

One then defines the matrix

$$A = D^{-1}W, \tag{7}$$

where D is a diagonal matrix with $d_{ii} = \sum_{j=1}^{L} w_{ij}$. A is then a Markov transition matrix for the graph defined by W. One can view nodes i and j of the graph as "similar" if a random walker starting at node i, and a random walker starting at node j, both evolving on the graph based on A, have similar probability distributions for their location after time t. The probability for a random walker starting at node i to arrive at node k at time t is given by A^t_{ik}; thus, to compare nodes i and j at time t, we should compare rows i and j of A^t. One could use the standard 2-norm, but vertices of higher degree would then contribute more to the overall metric. Therefore, one considers a relevant distance metric as:

$$D_t^2(i,j) = \sum_{k=1}^{L} \frac{\left(A^t_{ik} - A^t_{jk}\right)^2}{d_{kk}}. \tag{8}$$

This is the *diffusion distance* between data points i and j at time t. It can be shown that

$$D_t^2(i,j) = \sum_{k=0}^{L-1} \mu_k^{2t} (v_k(i) - v_k(j))^2, \tag{9}$$

where $v_0, \ldots, v_{L-1}$ and $\mu_0, \ldots, \mu_{L-1}$ are the eigenvectors and eigenvalues, respectively, of A. One orders the eigenvalues and eigenvectors so that $|\mu_0| \geq |\mu_1| \geq \cdots \geq |\mu_{L-1}|$. Since A is row-stochastic, $\mu_0 = 1$ and $v_0 = (1, \ldots, 1)^T$, so that the first term in the sum in Eq. 9 does not contribute to the diffusion distance. The diffusion map embedding is defined as $x_i \mapsto \left(\mu_1^t v_1(i), \mu_2^t v_2(i), \ldots, \mu_{L-1}^t v_{L-1}(i)\right)$; this embedding preserves the diffusion distance between data points. It is, however, not unusual to observe a *spectral gap* at l, such that $|\mu_0| \geq |\mu_1| \geq \cdots \geq |\mu_l| \gg |\mu_{l+1}| \geq \cdots \geq |\mu_{L-1}|$. Then, the computation of the diffusion distance can be truncated at l without significant loss of accuracy, and the *truncated* diffusion map embedding can be defined as $x_i \mapsto \left(\mu_1^t v_1(i), \mu_2^t v_2(i), \ldots, \mu_l^t v_l(i)\right)$. In our examples here we will use the $t = 0$ embedding.

DMAPs is supported by the theory of continuous operators on manifolds. In the limit of infinite data ($L \to \infty$), a random walk on the data, defined by the matrix A, converges to a random walk on a manifold Ω in continuous space (Nadler et al., 2006a,b). The matrix operator A can then be written as an integral operator on the manifold, $\overline{A}$,

$$\overline{A}v(x) = \int_\Omega \frac{k(x,y)}{d(x)} v(y) p(y) dy, \tag{10}$$

where $p(x)$ is the local density of the data and $d(x) = \int_\Omega k(x,y)p(y)dy$. It can be shown that (Coifman and Lafon, 2006)

$$\overline{A}v(x) = v(x) + \varepsilon\left(\frac{\Delta(vp)}{p} - v\frac{\Delta p}{p}\right) + O(\varepsilon^{3/2}), \tag{11}$$

so that

$$\lim_{\varepsilon \to 0} \frac{\overline{A} - I}{\varepsilon} v = \frac{\Delta(vp)}{p} - v\frac{\Delta p}{p}. \tag{12}$$

For stochastic problems which, on the manifold, are governed by a Langevin equation, $\dot{x} = -\nabla U + \sqrt{2}\dot{w}$ (w is a Brownian motion), the equilibrium density would be $p(x) = e^{-U(x)}$. If the data are sampled with this density, then

$$\lim_{\varepsilon \to 0} \frac{\overline{A} - I}{\varepsilon} v = \Delta v - 2\nabla v \cdot \nabla U. \tag{13}$$

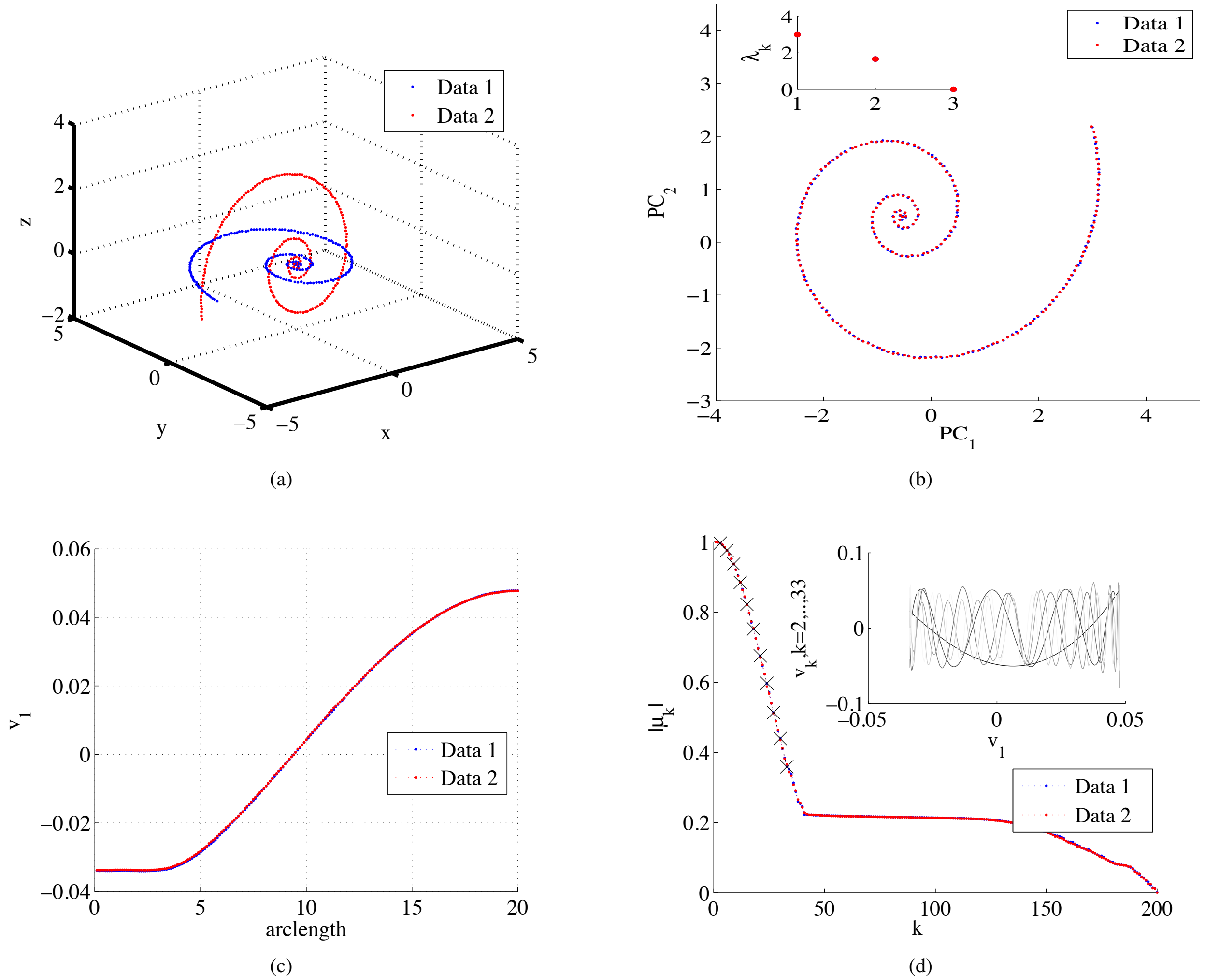
This operator is the *Fokker-Planck* operator with potential $2U$. Therefore, the eigenvectors of A approximate the eigenfunctions of this operator, and the eigenvalues of the Fokker-Planck operator, γ_k, are related to the eigenvalues of A by $\gamma_k = \frac{\mu_k - 1}{\varepsilon}$. Diffusion maps are, therefore, in such limiting cases, capable of recovering important features of the underlying stochastic dynamics.

Some Comparative Case Studies

Spiral

Our first example is designed to illustrate the different features of the three techniques above. We constructed a data set consisting of points evenly distributed along a planar spiral. We then embedded this spiral in three dimensions via two different embeddings. In the first, the spiral plane was offset slightly from the $x - y$ plane; in the second embedding, the slight offset was from the $x - z$ plane. In both cases Gaussian noise was added to the data points in all three directions (see Figure 1(a)).

The data set is effectively one-dimensional, as the points can be parameterized by the arclength along the spiral. However, the data is not well approximated by any one-dimensional linear subspace, but rather lie in an effectively two-dimensional linear subspace. Linear methods, such as PCA and LFA, are thus expected to perform differently than nonlinear ones (DMAPs).



(a)

(b)

(c)

(d)

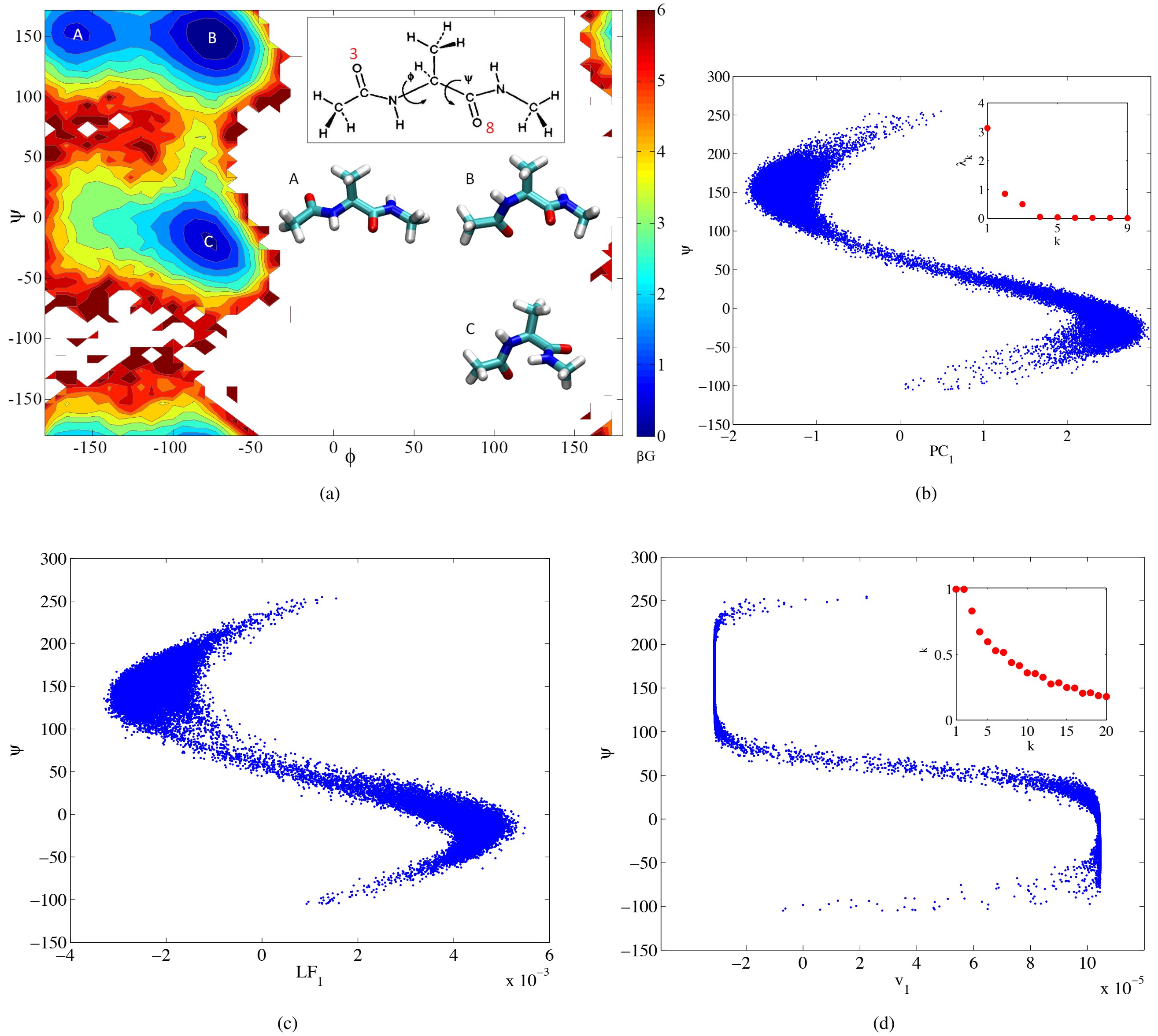
Figure 1: *(a) Two spiral data sets. (b) The data in principal component space. Inset: PCA eigenvalue spectrum. (c) Leading non-trivial DMAPs eigenvector recovers the arclength along the spiral. (d) DMAPs eigenvalue spectrum. Inset: Higher DMAPs eigenvectors, corresponding to the eigenvalues marked in Figure 1(d), plotted against the first non-trivial eigenvector. Note that they are well-correlated.*

Indeed, PCA successfully passes a plane through the data, thereby reducing the dimensionality from 3 to 2 (Figure 1(b)). LFA identifies the two "most representative" original directions for each data set (x and y in the case of data set 1, and x and z for data set 2). Since it is not constrained by linearity, DMAPs is able to reduce the data dimensionality from 3 to 1, and correctly uncovers the arclength as a single important variable for each spiral (Figure 1(c)). As seen in the inset of Figure 1(d), many successive leading eigenvectors from the Diffusion Map analysis are all correlated (they are higher harmonics of the first eigenvector, corresponding to the "arclength dimension"); other dimensions, transverse to the arclength, become manifest beyond the "knee" in the eigenvalue plot. We should at least briefly mention that many technical issues (such as the performance of the approach as the noise becomes stronger, or the optimal choice of ε) are still the subject of research.

Alanine Dipeptide

The second illustrative example we present involves the small biomolecule alanine dipeptide (N-acetyl-L-alanine-N-methylamide) (Ala2) (see Figure 2(a), inset). Ala2 has been the subject of numerous studies (e.g. Bolhuis et al. (2000); Hummer and Kevrekidis (2003); Ferguson et al. (2011)) and it is established that "good" physical variables which can parameterize its effective free energy surface are the backbone dihedral angles ψ [N-C_α-C-N] and ϕ [C-N-C_α-C]. We would like to explore whether the different data-mining methods are able to "uncover" these variables directly from simulation data.

A molecular dynamics simulation of Ala2 in explicit solvent was performed using the AMBER 10 molecular simulation package (Case et al., 2008) with an optimized version (Best and Hummer, 2009) of the AMBER ff03 force field (Duan et al., 2003). The simulation box contained 638 TIP3P water molecules (Jorgensen et al., 1983). We used periodic boundary conditions and the particle mesh Ewald



(a)

(b)

(c)

(d)

Figure 2: *(a) Ramachandran effective free energy surface of Ala2 and the structures corresponding to wells A, B, and C. Inset: Ala2 structure with dihedral angles ϕ and ψ and oxygen atoms 3 and 8 labeled. (b) Correspondence between leading PCA eigenvector and dihedral angle ψ. Inset: PCA eigenvalue spectrum. (c) Correspondence between leading Local Feature and dihedral angle ψ. (d) Correspondence between leading Diffusion Map eigenvector and dihedral angle ψ. Inset: DMAPs eigenvalue spectrum.*

method (Essmann et al., 1995) for long-range electrostatic interactions. The simulation was performed at constant volume. The temperature was maintained at 300K using a Langevin thermostat (Loncharich et al., 1992). Bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm (Ryckaert et al., 1977). The simulation used a time step of 0.001 ps. Data was collected for 7ns, with configurations saved every 0.1 ps. Figure 2(a) shows the resulting effective free energy surface $\beta G(\phi,\psi)$. We observe (as expected) two broad wells with three minima. The minima are located at $\phi \approx -160^\circ$, $\psi \approx 150^\circ$; $\phi \approx -75^\circ$, $\psi \approx 150^\circ$; and $\phi \approx -75^\circ$, $\psi \approx -20^\circ$. For the purposes of this discussion we label them A, B and C respectively (see Figure 2(a)).

The data used in our analysis consists of 50,000 consecutive snapshots from the molecular dynamics trajectory. Each data point is represented by a vector containing the physical coordinates of all atoms in Ala2 except the hydrogens. All snapshot configurations were aligned relative to a template to minimize the RMSD between them using the Kabsch algorithm (Kabsch, 1976, 1978).

PCA effectively reduces the dimensionality to three, and we see that the leading principal component is strongly correlated with the dihedral angle ψ (see Figure 2(b)). LFA is based on the same three eigenvectors as PCA and, based on our approach to selecting the first LFA seed, Figures 2(b) and 2(c) show similar results.

Three-dimensional embeddings of the data show that both

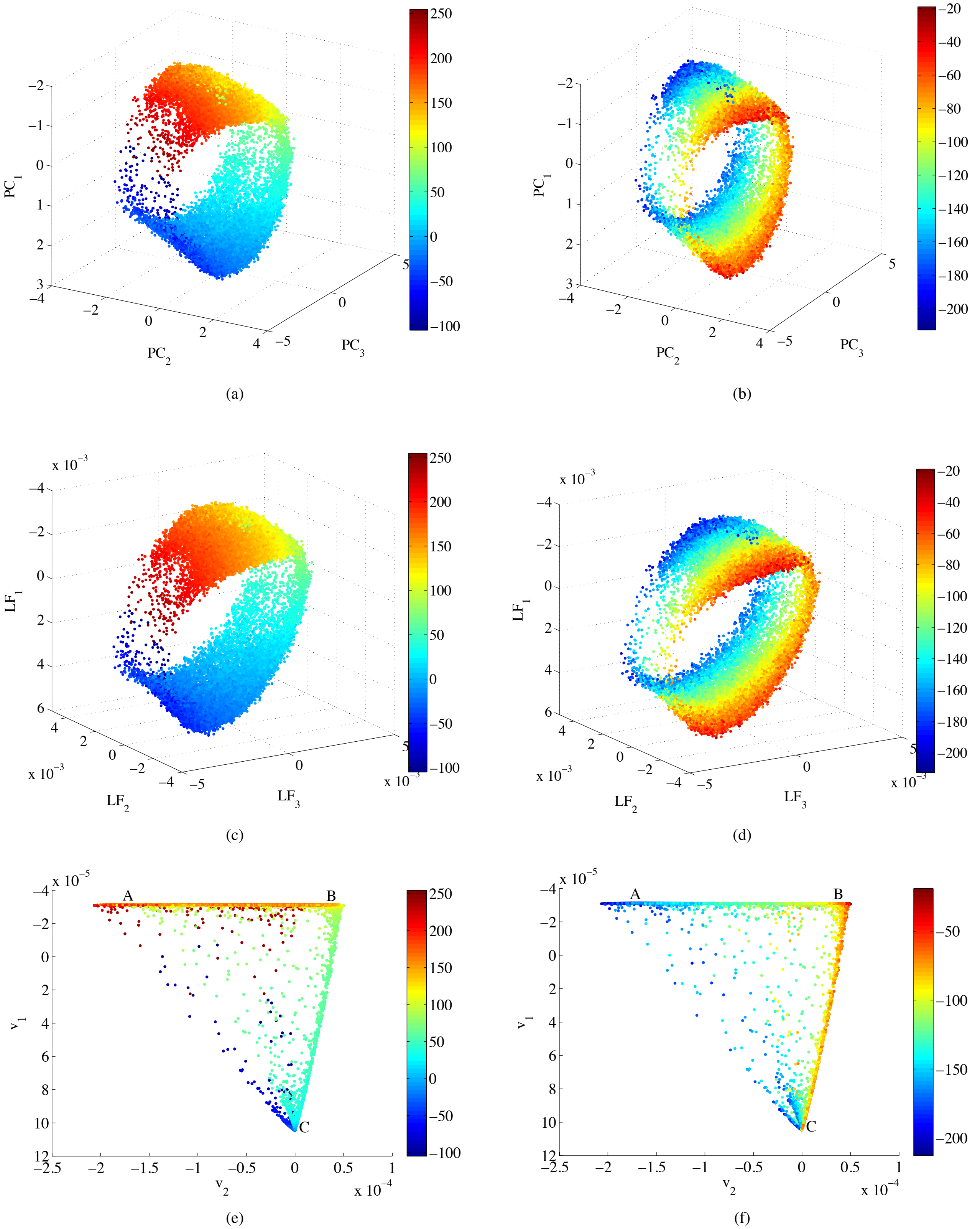


Figure 3: *(a),(b) 3D PCA view. (c), (d) 3D LFA view. (e), (f) Data in DMAPs space. (a), (c), (e) Data colored according to dihedral angle ψ. (b), (d), (f) Data colored according to dihedral angle ϕ.*

PCA and LFA organize the data well visually. Clearly, points are organized according to the values of the two dihedral angles, and in this way one can identify sections of the $\phi - \psi$ free energy surface (see Figures 3(c) and 3(d)). The data appear to approximately lie on an two-dimensional manifold in three-dimensional space. In Figure 3(c), it appears that a polar angle in the $LF_1 - LF_3$ plane is strongly correlated with ψ. Figure 3(d) shows a more direct correspondence between LF_2 and ϕ.

The corresponding PCA plots are shown in Figures 3(a) and 3(b) for completeness. The surface is now rotated and stretched, since the linear mappings from the original position data to the LFA and PCA coordinates are different. In both cases, it appears that by using *three components* (PC or LF), the values of ψ and ϕ can be simultaneously estimated.

It is natural to ask whether the values of ψ and ϕ can be estimated from only two reduced coordinates. In this example, LF_1 can be used to determine if the system is in the upper (red) or lower (blue) well in the free energy landscape plot (Figure 2(a)), but the information to distinguish red from green (close to the transition point) is no longer available without LF3. If one is only interested in which energy well the system resides, then LF_1 and LF_2 may provide a sufficient classification. However, if a model capable of describing the transition dynamics is desired, then one may wish to include LF_3.

Recall that LFA contains a topographic kernel, so that each coordinate in the reduced space is associated with a coordinate in the original, full space. The selection algorithm chose Atom 8 (direction x) for LF_1 since it has the highest variance, and subsequently selected Atom 3(y) as LF_2 and then Atom 8(z) as LF_3, since they were maximally uncorrelated with the previous local features. Atoms 3 and 8 are the two oxygen atoms in the molecule. By examining how the individual atoms move in configurations along the free energy landscape, we can rationalize why these two oxygen atoms are so significant. As the dihedral angle ϕ flips between the two wells, Atom 3 is maximally displaced, while a flip of angle ψ causes Atom 8 to flip relative to the central carbon.

An additional advantage of LFA is that, because each coordinate in the reduced space is associated with an atom's position, the time-evolution of these reduced coordinates can be understood as pertaining to the motion of the corresponding seed atom. In PCA, the time evolution of the principal components, which are global in nature, has no clear connection to the time evolution of the original variables.

As was the case with PCA, the leading non-trivial DMAPs eigenvector is strongly correlated with the angle ψ (see Figure 2(d)). Observing the data embedded in DMAPs space (Figures 3(e) and 3(f)), one perceives a structure resembling a triangle whose upper edge (where the approximate location of minima A and B are indicated) represents the entire upper free energy well. The bottom well is compressed into the lower vertex of the triangle, where the minimum C is located. Data points along the transition between minima B and C are found along the BC side of the triangle. One might argue that the *two-dimensional* projection using DMAPs provides a slightly clearer representation of the transitions than the corresponding two-dimensional PCA or LFA projections.

Conclusions

It seems clear to us, and -we hope- to the reader, even through such brief illustrations, that data-based model reduction is at the beginning of a new period of research and growth. The main issue we pointed out above is the choice of good *physical variables* corresponding to the (in some sense) optimal variables that PCA or DMAPs can locate - variables that may serve very well in computational tasks, but have no obvious physical meaning. LFA, through its choice of seeds, attempts to find “the best” physical variables that can be used to interpret the PCA based reduction. For DMAPs the corresponding step (finding good physical variables that are one-to-one with the DMAP-located variables) is a difficult and, at the moment, completely *ad hoc* post-processing task: one guesses possible candidate variables, and tests whether they are one-to-one with the selected diffusion map coordinates.

Whether with physically meaningful reduced sets of variables, or with useful -but physically not directly interpretable- such sets, the task of model reduction is only starting. Choosing the nature of the effective models in the new variables (deterministic or stochastic, discrete time or continuous time), determining how to obtain the effective equations in the new variables (off-line or on-line), using these variables to aid in biasing the computations to effectively search the system phase space, are all crucial tasks that are currently the subject of intense research (e.g. Ferguson et al. (2010)). Factoring out symmetries before (or as part of) data-mining is also an important part of the process.

We close by stating what, in our opinion, is the obvious and clear advantage of each method. Linear (PCA and LFA) methods provide a simple and systematic reconstruction (or *lifting*): constructing physical realizations consistent with the reduced variables; the corresponding task for DMAPs can be very difficult. On the other hand, nonlinear reduction methods can be much more parsimonious in reducing data such as the spirals of our artificial illustration - and going this extra mile may be crucial in the overall success of model reduction (Kevrekidis et al., 2004). It is interesting to state that *global* nonlinear reduction techniques like DMAPs can be combined with *local* uses of linear tools (local PCA) in model reduction studies, and that is also an open current research direction.

Acknowledgements

The work by Y.X., P.J.L., and M.A.G. was supported by ACS-PRF Grant # 48919-ND9. The work of L.V.N. and I.G.K. was partially supported by the US Department of Energy grant DE-SC0002097. C.J.D was supported by US Department of Energy grant DE-FG02-97ER25308.

References

Balsera, M., Wriggers, W., Oono, Y., Schulten, K. (1996). Principal Component Analysis and Long Time Protein Dynamics. *Journal of Physical Chemistry*, *100*, 2567–2572.

Best, R. B., Hummer, G. (2009). Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *Journal of Physical Chemistry B*, *113*, 26, 9004–9015.

Bolhuis, P. G., Dellago, C., Chandler, D. (2000). Reaction Coordinates of Biomolecular Isomerization. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11, 5877–5882.

Case, D. A., Darden, T. A., T. E. Cheatham, I., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossvry, I., Wong, K. F., Paesani, F., Vanicek, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Kollman, P. (2008). AMBER 10. University of California, San Francisco.

Coifman, R. R., Lafon, S. (2006). Diffusion Maps. *Applied and Computational Harmonic Analysis*, *21*, 1, 6–31.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., Zucker, S. W. (2005a). 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.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., Zucker, S. W. (2005b). 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.

Daoud, M., De Gennes, P. (1979). Some Remarks on the Dynamics of Polymer Melts. *Journal of Polymer Science: Polymer Physics Edition*, *17*, 11, 1971–1981.

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G. M., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. M., Kollman, P. (2003). A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *Journal of Computational Chemistry*, *24*, 16, 1999–2012.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., Pedersen, L. G. (1995). A Smooth Particle Mesh Ewald Method. *Journal of Chemical Physics*, *103*, 19, 8577–8593.

Ferguson, A. L., Panagiotopoulos, A. Z., Debenedetti, P. G., Kevrekidis, I. G. (2010). 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, 13597–13602.

Ferguson, A. L., Panagiotopoulos, A. Z., Debenedetti, P. G., Kevrekidis, I. G. (2011). Integrating Diffusion Maps with Umbrella Sampling: Application to Alanine Dipeptide. *Journal of Chemical Physics*, *134*, 13.

Hummer, G., Kevrekidis, I. G. (2003). Coarse Molecular Dynamics of a Peptide Fragment: Free Energy, Kinetics, and Long-Time Dynamics Computations. *Journal of Chemical Physics*, *118*, 23, 10762–10773.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., Klein, M. L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics*, *79*, 2, 926–935.

Kabsch, W. (1976). Solution for Best Rotation to Relate 2 Sets of Vectors. *Acta Crystallographica Section A*, *32*, 922–923.

Kabsch, W. (1978). Discussion of Solution for Best Rotation to Relate 2 Sets of Vectors. *Acta Crystallographica Section A*, *34*, 827–828.

Kevrekidis, I. G., Gear, C. W., Hummer, G. (2004). Equation-Free: The Computer-Aided Analysis of Complex Multiscale Systems. *AIChE Journal*, *50*, 7.

Loncharich, R. J., Brooks, B. R., Pastor, R. W. (1992). Langevin Dynamics of Peptides - the Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-Methylamide. *Biopolymers*, *32*, 5, 523–535.

Nadler, B., Lafon, S., Coifman, R. R., Kevrekidis, I. G. (2006a). Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems. *Applied and Computational Harmonic Analysis*, *21*, 1, 113–127.

Nadler, B., Lafon, S., Kevrekidis, I. G., Coifman, R. R. (2006b). Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators. *Advances in Neural Information Processing Systems*, *18*, 955–962.

Penev, P. S., Atick, J. J. (1996). Local Feature Analysis: A General Statistical Theory for Object representation. *Computation in Neural Systems*, *7*, 477–500.

Ryckaert, J. P., Ciccotti, G., Berendsen, H. J. C. (1977). Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *Journal of Computational Physics*, *23*, 3, 327–341.

Shlens, J. (2005). A Tutorial on Principal Component Analysis.

Xue, Y., Ludovice, P. J., Grover, M. A. (2010). Local Feature Analysis based Clustering Algorithm with Application to Polymer Dynamics Model Reduction. In *Proceedings of the 49th IEEE Conference on Decision and Control*, pp. 1687–1692.

Xue, Y., Ludovice, P. J., Grover, M. A. (in preparation). Dynamics Coarse Graining for Complex System, from Theory to Simulation.

Zhang, Z., Wriggers, W. (2008). Coarse-Graining Protein Structures With Local Multivariate Features from Molecular Dynamics. *Journal of Physical Chemistry, B*, *112*, 14026–14035.