Geometric and Markov Chain Based Methods for Data Analysis

Danielle Middlebrooks

Advisors:
Maria Cameron Kasso Okoudjou

Applied Mathematics, Statistics and Scientific Computation
University of Maryland- College Park

January 22, 2017

1 Introduction

2 Part 1

- Diffusion Maps
- Transition Path Theory
- Allostery of BirA Protein

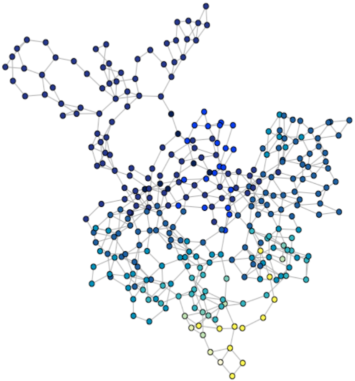
3 Part 2

- Point Cloud Discretization for Committor Functions
- Multiscale Geometric Methods

4 Conclusion

Introduction

Wild type BirA Protein



- The use of networks is a popular tool for data representation and organization.
- In general, a network consist of a set of nodes along with a set of edges.
- New tools need to be developed for the analysis of complex networks.
- Particularly interested in applications arising from chemical physics.

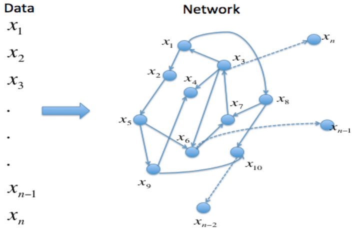
Motivation

Suppose a protein has two metastable states. We are interested in a way to quantify the transition between the two states.

Motivation

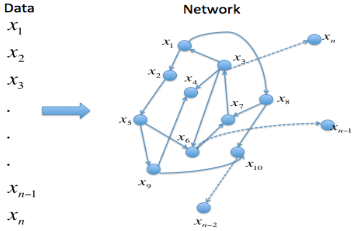
Suppose a protein has two metastable states. We are interested in a way to quantify the transition between the two states.

Approach:



Motivation

Suppose a protein has two metastable states. We are interested in a way to quantify the transition between the two states. Approach:



Goal

Develop computational techniques for data analysis and the study of transition processes between metastable sets of states.

Outline

Discuss existing computational tools for network analysis as well as describe ongoing projects.

Part 1:

- Diffusion Maps (Coifman and Lafon, 2006)
- Transition Path Theory (E and Vanden-Eijnden, 2006)
- The study of allostery in BirA protein (with S. Matysiak and G. Custer)

Outline

Discuss existing computational tools for network analysis as well as describe ongoing projects.

Part 1:

- Diffusion Maps (Coifman and Lafon, 2006)
- Transition Path Theory (E and Vanden-Eijnden, 2006)
- The study of allostery in BirA protein (with S. Matysiak and G. Custer)

Part 2:

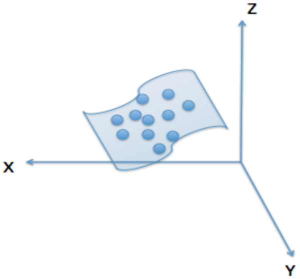
- Computation of the committor function on point clouds (Lai and Lu, 2017)
- Multiscale Geometric Methods (Maggioni and Collaborators, 2011)

1. Introduction
2. Part 1
   - Diffusion Maps
   - Transition Path Theory
   - Allostery of BirA Protein
3. Part 2
   - Point Cloud Discretization for Committor Functions
   - Multiscale Geometric Methods
4. Conclusion

Motivation

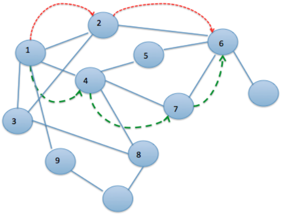
Methodology for finding hidden geometric structure in large high-dimensional data sets.

Idea: The diffusion map embeds the data into Euclidean space of lower dimension so that the Euclidean distance is equal to the diffusion distance in original space.



Motivation

Consider a discrete time Markov chain. Suppose we take a random walk on our data, jumping between data points.



The probability of jumping between two data points, x_i and x_j in one step of the random walk is defined by the stochastic matrix $P_{ij} = p(x_i, x_j)$.

The probability of transition from x_i to x_j in t jumps is given by $p_t(x_i, x_j)$ or P^t. The idea is taking larger powers of P will allow us to integrate the local geometry and thus reveal relevant geometric structures of the dataset at different scales.

Diffusion Distance to Diffusion Map

First, define a positive symmetric kernel $k(x_i, x_j) = h(\|x_i - x_j\|/\epsilon)$ with scaling parameter ϵ. Often, this kernel is of the form

$$k(x_i, x_j) = e^{\frac{-\|x_i - x_j\|^2}{\epsilon}} \tag{1}$$

This kernel is converted into a stochastic matrix P. The diffusion distance is defined by

$$D_t(x_i, x_j)^2 = ||p_t(x_i, \cdot) - p_t(x_j, \cdot)||^2_{L^2}, \quad t = 1, 2, 3 \dots \tag{2}$$

where $p_t(x_i, \cdot)$ are the rows of the matrix P^t.

Diffusion Distance to Diffusion Map

Up to some relative precision δ, we have

$$D_t(x_i, x_j)^2 = \sum_{l=1}^{s} \lambda_l^{2t} (\psi_l(x_i) - \psi_l(x_j))^2 \quad \text{with } s = s(\delta, t) \qquad (3)$$

where λ_l, ψ_l are the are the eigenvalues and eigenvectors the matrix P. Thus the family of diffusion maps $\{\Psi_t\}_{t\in\mathbb{N}}$ is given by

$$x_i \mapsto \Psi_t(x_i) = \begin{bmatrix} \lambda_1^t \psi_1(x_i) \\ \lambda_2^t \psi_2(x_i) \\ \vdots \\ \lambda_s^t \psi_s(x_i) \end{bmatrix} \qquad (4)$$

The components of $\Psi_t(x_i)$ are the diffusion coordinates of x_i. The map Ψ_t embeds the data set into a s-dimensional Euclidean space.

Key Points

- Diffusion maps method is robust to noise perturbation.
- Allows for geometric analysis at different scales.
- Accuracy of results is heavily determined by user-defined parameter ϵ

1 Introduction

2 Part 1

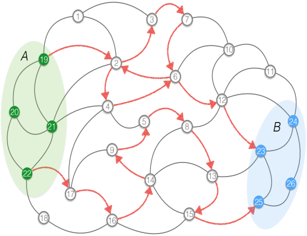
- Diffusion Maps
- Transition Path Theory
- Allostery of BirA Protein

3 Part 2

- Point Cloud Discretization for Committor Functions
- Multiscale Geometric Methods

4 Conclusion

Markov Jump Processes



Consider a Markov jump process on discrete state space S with infinitesimal generator $L = (L_{i,j})_{i,j \in S}$:

$$\begin{cases} L_{i,j} \geq 0, & \forall i,j \in S, i \neq j \\ L_{i,i} = -\sum_{j \neq i} L_{i,j}, & \forall i \in S \end{cases}$$

Assume L is irreducible and the MJP is ergodic with respect to equilibrium probability distribution $\pi = (\pi_i)_{i \in S}$ which satisfies

$$\pi^T L = 0$$

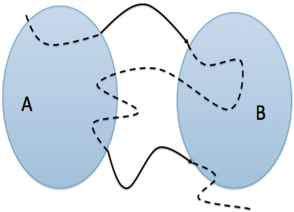
$$\sum_{i \in S} \pi_i = 1$$

What is Transition Path Theory?

Transition Path Theory (TPT) is a framework to analyze the statistical properties of reactive trajectories in which transitions occur between two disjoint subsets

TPT can be applied to

- Physics
- Chemistry
- Biology
- Social Sciences



Key Concepts of Transition Path Theory

For simplicity we consider the time reversible case.

The committor function $q = (q_i)_{i \in S}$ is the probability that starting at a state i, the trajectory will reach set B prior to set A and satisfies

$$\begin{cases} \sum_{j \in S} L_{i,j} q_j = 0, & \text{if } i \in S \setminus (A \cup B) \\ q_i = 0, & \text{if } i \in A \\ q_i = 1, & \text{if } i \in B \end{cases}$$

Key Concepts of Transition Path Theory

For simplicity we consider the time reversible case.

The committor function $q = (q_i)_{i \in S}$ is the probability that starting at a state i, the trajectory will reach set B prior to set A and satisfies

$$\begin{cases} \sum_{j \in S} L_{i,j} q_j = 0, & \text{if } i \in S \setminus (A \cup B) \\ q_i = 0, & \text{if } i \in A \\ q_i = 1, & \text{if } i \in B \end{cases}$$

Probability current of reactive trajectories is given by

$$f_{i,j} = \begin{cases} (1 - q_i)\pi_i L_{i,j} q_j, & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

Key Concepts of Transition Path Theory

For simplicity we consider the time reversible case.

The committor function $q = (q_i)_{i\in S}$ is the probability that starting at a state i, the trajectory will reach set B prior to set A and satisfies

$$\begin{cases} \sum_{j\in S} L_{i,j} q_j = 0, & \text{if } i \in S \setminus (A \cup B) \\ q_i = 0, & \text{if } i \in A \\ q_i = 1, & \text{if } i \in B \end{cases}$$

Probability current of reactive trajectories is given by

$$f_{i,j} = \begin{cases} (1 - q_i)\pi_i L_{i,j} q_j, & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

Reactive current

$$F_{i,j} = f_{i,j} - f_{j,i} = \pi_i L_{i,j}(q_i - q_j)$$

Key Concepts of Transition Path Theory

For simplicity we consider the time reversible case.

The committor function $q = (q_i)_{i \in S}$ is the probability that starting at a state i, the trajectory will reach set B prior to set A and satisfies

$$\begin{cases} \sum_{j \in S} L_{i,j} q_j = 0, & \text{if } i \in S \setminus (A \cup B) \\ q_i = 0, & \text{if } i \in A \\ q_i = 1, & \text{if } i \in B \end{cases}$$

Probability current of reactive trajectories is given by

$$f_{i,j} = \begin{cases} (1 - q_i)\pi_i L_{i,j} q_j, & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

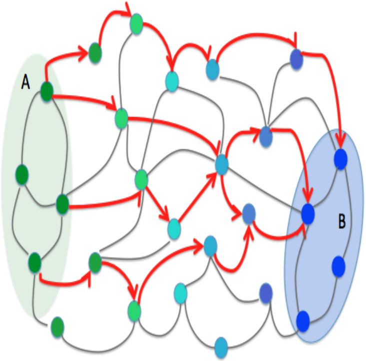
Reactive current

$$F_{i,j} = f_{i,j} - f_{j,i} = \pi_i L_{i,j}(q_i - q_j)$$

Transition rate

$$\nu_{AB} = \sum_{(i,j) \in cut} F_{ij}$$

Example



Examples of reactive trajectories are shown by red arrows.

Values of the committor are expressed by color. $q = 0$ are green, $q = 1$ are blue. Sequence of values of the committor strictly increase along reactive trajectories.

Key Points

- TPT gives quantitative characteristics of transition processes in Markov chains between any two disjoint subsets of states.
- Can find rates of transition even in Markov chains which are time-irreversible.
- Limitations lie in the ability to solve the system for the committor function.

1 Introduction

2 Part 1

- Diffusion Maps
- Transition Path Theory
- Allostery of BirA Protein

3 Part 2

- Point Cloud Discretization for Committor Functions
- Multiscale Geometric Methods

4 Conclusion

Analysis of Allostery in BirA Protein

Collaboration with Prof. S. Matysiak and G. Custer.

Goal: Study the information transfer between two regions in the BirA protein.

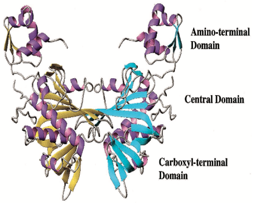


Figure: BirA Protein. Wood, Zachary A., et al.

Consider two sets:

- Ligand Binding Region (residues 212-223 and 116-124)
- Dimer-Interface region (residues 193-196 and 140-146)

Calculating Edge Weights

Idea: Model the BirA protein using a network and determine signal propagation between the two regions using TPT.

Calculating Edge Weights

Idea: Model the BirA protein using a network and determine signal propagation between the two regions using TPT.

Ribeiro et al. [7] would define edge weights as

$$\omega_{ij} = \begin{cases} \omega_b, & \text{if i and j are covalently bound} \\ \chi_{ij}, & \text{otherwise} \end{cases}$$

With ω_b a predefined value and χ_{ij} calculated by average interaction energy between all pairs of noncovalently bound residues. $\chi_{ij} \equiv 0.5\{1 - (\varepsilon_{ij} - \varepsilon_{av})/5\varepsilon_{rmsd}\}$

Calculating Edge Weights

Idea: Model the BirA protein using a network and determine signal propagation between the two regions using TPT.

Ribeiro et al. [7] would define edge weights as

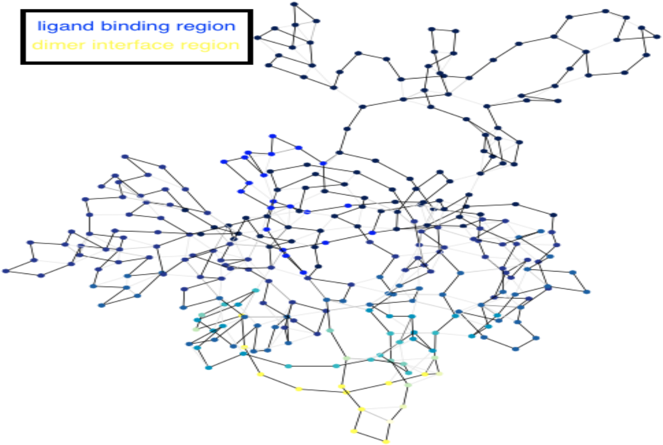
$$\omega_{ij} = \begin{cases} \omega_b, & \text{if i and j are covalently bound} \\ \chi_{ij}, & \text{otherwise} \end{cases}$$

With ω_b a predefined value and χ_{ij} calculated by average interaction energy between all pairs of noncovalently bound residues. $\chi_{ij} \equiv 0.5\{1 - (\varepsilon_{ij} - \varepsilon_{av})/5\varepsilon_{rmsd}\}$

For our analysis, the weights were calculated by taking the absolute value of interaction energy between pairs of residues.

Committor Function via Transition Path Theory

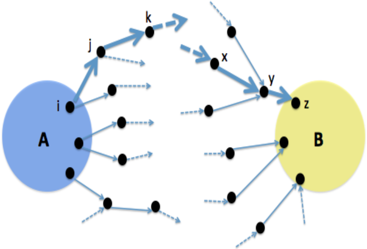
Wild-type apo BirA protein



Finding Pathways from A to B

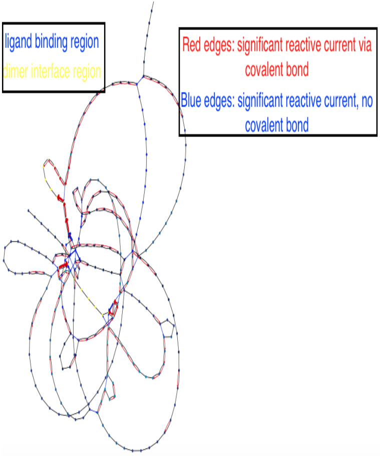
Algorithm (Finding Pathways)

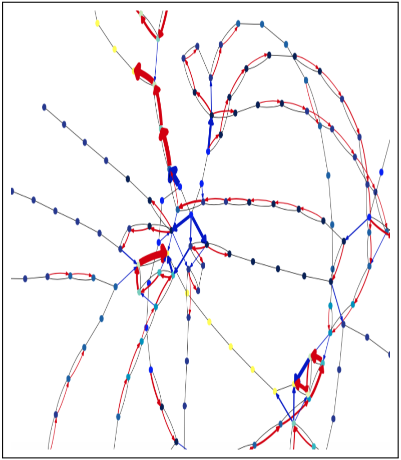
```
For all nodes $i$ in $A$
    sort edges $i \to j$ according to the current they carry
    find all $j_{max} >$ min threshold and add $i$ and $j_{max}$ to path
    print_path(path,$j_{max}$)
        sort edges $j_{max} \to k$ according to the current they carry
        find all $k_{max} >$ min threshold and add $k_{max}$ to the path
        print_path(path,$k_{max}$)
    continue until path reaches set $B$ or current is less than min threshold
```



Results for Reactive Current

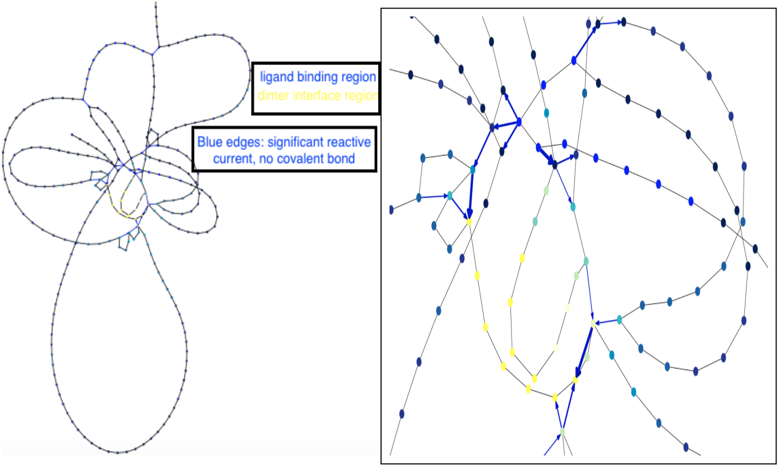
Wild-type apo BirA protein


ligand binding region
dimer interface region
Red edges: significant reactive current via covalent bond
Blue edges: significant reactive current, no covalent bond



Results for Reactive Current

Wild-type apo BirA protein without covalent bonds



ligand binding region

dimer interface region

Blue edges: significant reactive current, no covalent bond

Questions for Further Analysis

- What is the best way to define the weights of edges?
- How do mutations affect energy process?
- Should covalent bonds be taken into account?

1 Introduction

2 Part 1

- Diffusion Maps
- Transition Path Theory
- Allostery of BirA Protein

3 Part 2

- Point Cloud Discretization for Committor Functions
- Multiscale Geometric Methods

4 Conclusion

Point Cloud Discretization of q

Consider stochastic process of the overdamped Langevin equation

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dW_t$$

Again, q solves the following PDE with Dirichlet boundary conditions given on A and B:

$$\begin{cases} Lq = 0, & \text{in } \Omega \setminus (A \cup B) \\ q = 0, & \text{in A }, \\ q = 1 & \text{in B} \end{cases}$$

with

$$L = -\beta^{-1}\Delta + \nabla U \cdot \nabla$$

With transition rate

$$\nu_{AB} = \int_\Sigma J_R(x) \cdot \hat{n}(x) d\sigma(x)$$

and $J_R = Z^{-1}exp(-\beta U)\nabla q$

Motivation

Solving the PDE is non-trivial due to the curse of dimensionality.

Motivation

Solving the PDE is non-trivial due to the curse of dimensionality.

Main Idea

Transition between regions A and B lie in low-dimensional manifold and thus can approximate q on point cloud rather than entire configuration space.

Motivation

Solving the PDE is non-trivial due to the curse of dimensionality.

Main Idea

Transition between regions A and B lie in low-dimensional manifold and thus can approximate q on point cloud rather than entire configuration space.

Goal

Directly solve the committor equation based on point cloud discretization (Local Mesh Method).

Local Mesh Method

Rongjie Lai and Jianfeng Lu

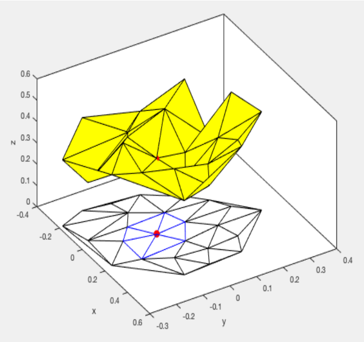
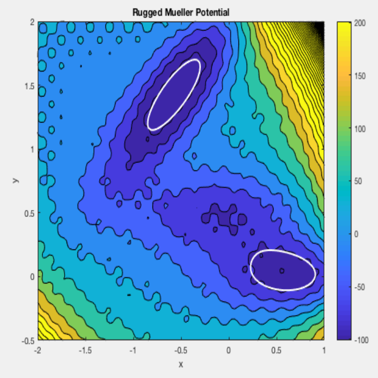


Figure: Projection of KNN with first ring

- At a point p_i project its K - nearest neighborhood on the tangent plane of p_i.
- Approximate stiffness matrix: $A = \sum_{F\in R(i)} \int_F e^{-\beta U} \nabla_F \eta_i \cdot \nabla_F \eta_j$
- Symmetrized stiffness matrix: $S = \begin{cases} \max\,(A_{ij}, A_{ji}) & \text{if } A_{ij} \leq 0 \text{ and } A_{ji} \\ \min\,(A_{ij}, A_{ji}) & \text{if } A_{ij} \geq 0 \text{ and } A_{ji} \\ \min\,(A_{ij}, A_{ji}) & \text{if } A_{ij} \cdot A_{ji} < 0 \\ -\sum_{k\neq i} S_{ik} & \text{if } i = j \end{cases}$

Rugged Mueller Potential Function



- Potential $U_{rm}(x, y) = U + \gamma \sin(2k\pi x) \sin(2k\pi y)$ with U orginal Mueller potential
- Consider sets $A = \{U(x, y) < -120\} \cap \{y > 0.75\}$ and $B = \{U(x, y) < -82\} \cap \{y < 0.35\}$

Committor Function and Reactive Current

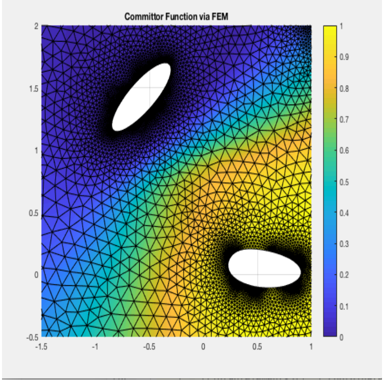


Figure: Committor function via FEM

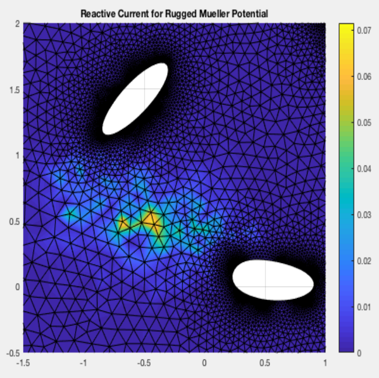


Figure: Reactive current

Results of Committor Function

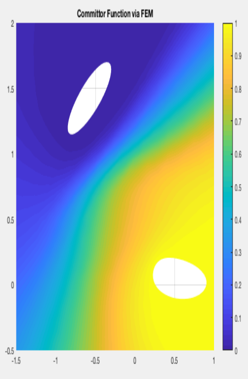


Figure: Finite Element Method

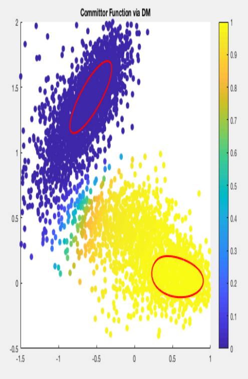


Figure: Diffusion Maps Method

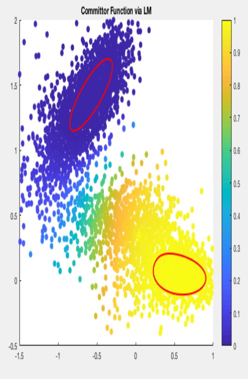


Figure: Local Mesh Method

Comparison of Results

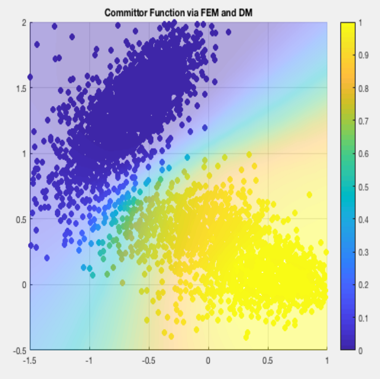


Figure: Committor function via FEM and DM

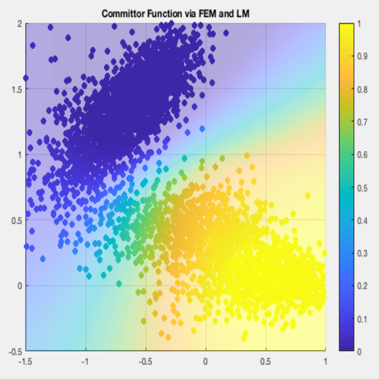
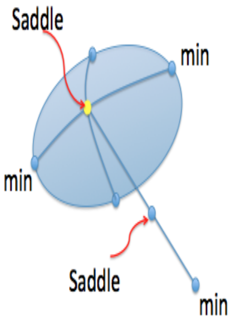


Figure: Committor function via FEM and LM

“Umbrella” Potential Function



- Potential $U = 10u^2(u^2-1) + 0.07x^2 + y^2 + \frac{15}{19\sqrt{(19)}}z^2$ with $u = \frac{x^2+z^2}{4} + y$
- Consider sets containing two local minima along canopy of umbrella

Result of Committor Function

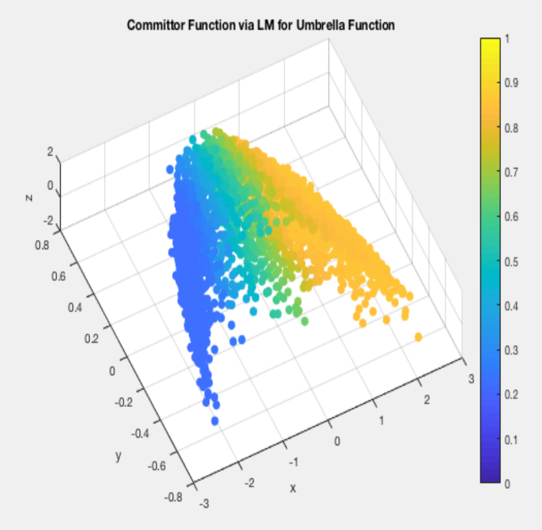


Figure: Committor function via Local Mesh method

Key Points

- The local mesh method provides a tool to analyze the stochastic system in the framework of TPT.
- This method is sensitive to the the number of points used.
- We would like to first learn the manifold in order to get a better arrangement of points.

1 Introduction

2 Part 1

- Diffusion Maps
- Transition Path Theory
- Allostery of BirA Protein

3 Part 2

- Point Cloud Discretization for Committor Functions
- Multiscale Geometric Methods

4 Conclusion

Motivation: Principal Component Analysis

One main assumption of these high-dimensional data sets is they are often modeled as low-dimensional sets embedded in high dimensional space.

Question: How do we estimate the intrinsic dimension of high dimensional datasets?

Motivation: Principal Component Analysis

One main assumption of these high-dimensional data sets is they are often modeled as low-dimensional sets embedded in high dimensional space.

Question: How do we estimate the intrinsic dimension of high dimensional datasets?

Represent data matrix $X \in \mathbb{R}^{Dxn}$. Can compute Singular Value Decomposition

$$X = U\Sigma V^T$$

U: Orthogonal D x D matrix
Σ: diagonal D x n matrix of singular values
V: Orthogonal n x n matrix

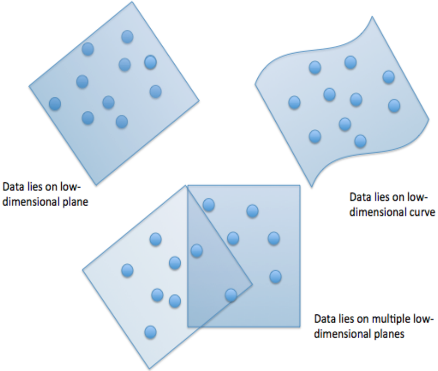
Motivation: Principal Component Analysis

If data lies on k-dimensional hyperplane, only k singular values are nonzero and the first k columns of U span the desired hyperplane.

Motivation: Principal Component Analysis

If data lies on k-dimensional hyperplane, only k singular values are nonzero and the first k columns of U span the desired hyperplane.

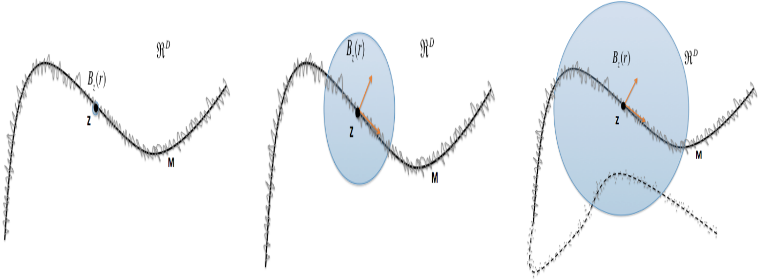
What if the data is non-linear?


Data lies on low-dimensional plane
Data lies on low-dimensional curve
Data lies on multiple low-dimensional planes

Multiscale SVD (MSVD)

Problem Description

Consider data $\{x_i\}_{i=1}^n$ sampled from a manifold of dimension k embeded in $\mathbb{R}^D$ with $k << D$. Our observations will be $\{\bar{x_i}\}_{i=1}^n$ with $\bar{x}_i = x_i + \sigma\eta_i$ and $\sigma\eta_i$ are drawn i.i.d. from noise random variable N.



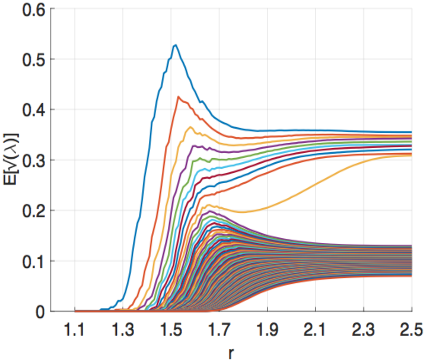
Algorithm for Estimating Intrinsic Dimension

Algorithm (MSVD)

```
Set noisy data: $\tilde{X}$      Fix: $\tilde{z} \in \{\tilde{x}_1, ..., \tilde{x}_n\}$
For each $\tilde{z}$, $r > 0$, $i = 1, ..., D$
    Compute: $\tilde{X}_{n,\tilde{z},r} = \tilde{X}_n \cap B_{\tilde{z}}(r)$
             $cov(\tilde{X}_{n,\tilde{z},r}) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \mathbb{E}_n[\tilde{X}]) \otimes (\tilde{x}_i - \mathbb{E}_n[\tilde{X}])$
             $\lambda_i^{(\tilde{z},r)} = \lambda_i(cov(\tilde{X}_{n,\tilde{z},r}))$
Observe $\tilde{k}$ large eigenvalues and $D - \tilde{k}$ smaller noise eigenvalues.
$R_{min} \leftarrow$ smallest scale for which $(\lambda_{\tilde{k}+1}^{(\tilde{z},r)})^{1/2}$ is decreasing.
$R_{max} \leftarrow$ largest scale $> R_{min}$ for which $(\lambda_{\tilde{k}+1}^{(\tilde{z},r)})^{1/2}$ is nondecreasing.
$k \leftarrow$ largest $i$ such that
    $r \in (R_{min}, R_{max})$, $(\lambda_i^{(\tilde{z},r)})^{1/2}$ is linear and $(\lambda_{i+1}^{(\tilde{z},r)})^{1/2}$ is quadratic in $r$
```

Example

$S^9(1000, 100, 0.1)$: 1000 points uniformly sampled on a 9-dimensional unit sphere, embeded in 100 dimensions with Gaussian noise of standard deviation 0.1 in every direction.



Key Points

- While PCA fails due to non-linearity of the data, this method introduces a multiscale geometric approach to estimating the intrinsic dimension of nonlinear data.
- Able to distinguish between the underlying k- dimensional structure of the data from the effects of noise and curvature.
- Depending on the dataset, finding a “good” range of radii may be difficult.

1 Introduction

2 Part 1
- Diffusion Maps
- Transition Path Theory
- Allostery of BirA Protein

3 Part 2
- Point Cloud Discretization for Committor Functions
- Multiscale Geometric Methods

4 Conclusion

Conclusion

- The use of networks to understand complex systems is increasingly becoming popular to analyze systems in a variety of fields.
- In many applications, it is of interest to study the transitions that occur between two groups of nodes. This can be analyzed by popular methods such as diffusion maps and transition path theory.
- Often, one can take advantage of the fact that complex high dimensional systems often have a low-lying intrinsic dimension. PCA and more recently MSVD are methods that help use determine these dimensions.
- While these methods are extremely popular, new tools still need to be developed for high dimensional data analysis.

Research Plans

- Continue the analysis of the allostery of BirA protein.
- Develop method for learning manifold from point clouds.
- Compute other functions on point clouds by solving other types of PDEs.

References

1 Weinan, E., and Vanden-Eijnden, Eric. “Towards a Theory of Transition Paths.” Journal of statistical physics 123.3 (2006).

2 Coifman, Ronald R., and Lafon, Step̀hane, “Diffusion maps.” Applied and computational harmonic analysis 21.1 (2006): 5-30

3 Little, Anna V., Maggioni, Mauro, and Rosasco, Lorenzo. “Multiscale geometric methods for estimating intrinsic dimension.” Proc. SampTA 4.2 (2011).

4 Lai, Rongjie, and Lu, Jianfeng. “Point cloud discretization of Fokker-Planck operators for committor functions.” arXiv preprint arXiv:1703.09359 (2017).

5 Kim, Sang Beom, et al. “Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein.” The Journal of chemical physics 142.8 (2015): $02B613_1$.

6 Metzner, Philipp, Schütte, Christof, and Vanden-Eijnden, Eric. “Transition path theory for Markov jump processes.” Multiscale Modeling & Simulation 7.3 (2009): 1192-1219

7 Ribeiro, Andre AST, and Ortiz, Vanessa. “Determination of signaling pathways in proteins through network theory: importance of the topology.” Journal of chemical theory and computation 10.4 (2014): 1762-1769.

Acknowledgements

- Thank you to my advisors and committee members
- National Science Foundation COMBINE-NRT Program under Grant No. DGE-1632976