Problems
Geometric and Graph methods for High-Dimensional Data

Wenjing Liao, Mauro Maggioni

Link to code: `http://www.math.jhu.edu/~mauro/code.html`

Session I

1. Concentration of measure phenomena, curse of dimensionality. Compute $\mathbb{E}[||x||^2]$ when $x \sim \mathcal{N}(0, I_D)$. Show that with overwhelming probability $||x||^2$ is very close to this expected value. What does this say about the geometry of a high-dimensional Gaussian random variable?

2. Formalize the informal reasoning to show that a histogram density estimator for a Lipschitz density on the unit cube in $\mathbb{R}^D$ needs at least ϵ^{-D} samples in order to achieve error ϵ. Can you be more precise about the power of ϵ that is in fact needed?

3. Multiscale SVD. Let $\mathcal{M}$ be a d-dimensional smooth compact manifold with no boundary. Pick a point $z \in \mathcal{M}$. Study the behavior of the covariance matrix of the volume measure ρ on $\mathcal{M}$ restricted to $B_r(z) \cap \mathcal{M}$.

Hint: do first the case where $\mathcal{M}$ has co-dimension 1. Represent locally the manifold as a graph of a function from T_z, its d-dimensional tangent plane at z, to $T_z^{\perp} \cong \mathbb{R}$, expand using Taylor and compute the covariance. Extend to the case of larger co-dimension, but do not expect nice formulas in this case (at least, as far as I know there are none).

4. Spectral graph theory: Let $G = (V, E)$ be an undirected graph with vertex set V. Each edge between two vertices x and y carries a non-negative weight $w(x, y)$. The *weighted adjacency matrix* of the graph is the matrix $W = \{w(x, y)\}_{x,y \in V}$. We consider undirected graph and require $w(x, y) = w(y, x)$. The degree of a vertex $x \in V$ is defined as $d_x = \sum_{y \in V} w(X, Y)$. The *degree matrix* $D = \text{diag}(d_x)_{x \in V}$.

Graph Laplacian $L = D - W$. Reason the following statements:

a. For every $f : V \to R$, $f^T L f = \frac{1}{2} \sum_{x,y \in V} w(x, y)[f(x) - f(y)]^2$. Hence L is positive definite. The smallest eigenvalue of L is 0 and the corresponding eigenvector is $\mathbf{1}_V$. The multiplicity k of the eigenvalue 0 equals to the number of connected components $\mathcal{V}_1, \ldots, \mathcal{V}_k$ in the graph. The eigen-space of eigenvalue 0 is spanned by the indicator function $\mathbf{1}_{\mathcal{V}_1}, \ldots, \mathbf{1}_{\mathcal{V}_k}$.

b. The eigenvalues of L is $0 = \lambda_0 \leq \lambda_1 \leq \ldots$. and the eigenvector corresponding to λ_1 satisfies

$$\min_{f : V \to R} f^T L f \quad \text{subject to } f \perp \mathbf{1}_V,\ \|f\| = 1.$$

5. Diffusion Geometry. Consider the space with latent variable y taking values in $\{0, 1\}$ with equal probabilities, and $x \in \mathbb{R}^D$ with $x|y = 0 \sim \mathcal{N}(0, \frac{1}{D} I_D)$ and $x|y = 1 \sim \mathcal{N}(2\mathbf{e}_1, \frac{1}{D} I_D)$, where $\mathbf{e}_1 = (1, 0, 0, \ldots, 0)$. Sample n times from x to obtain a data set X_n. Connect every point in X_n with all the points in a ball of radius ϵ. How do you expect the random walk on this proximity graph to look like? Do you expect spectral clustering to work? Download Diffusion Geometry code and run Demo_Clustering.m, Demo_SprayPaintedData.m and RunExamples.m. Discuss eigenvectors, localization, spectral clustering.

Session II

1. Principal Component Analysis, Covariance Matrices

a. Why are the principal components of the data set obtained by randomly drawing disks of a certain size on the unit square shown in the lecture equal to the Fourier basis? Any caveats? Does the answer depend on the shape (disk vs., say, square?)? Do the singular values depend on the shape?

b. Let $x \sim \mathcal{N}(0, I_D)$. How many samples do you think are needed in order for the empirical covariance matrix to be ϵ-close to the true covariance matrix? What if $x \sim \mathcal{N}(0, \Sigma)$ for Σ of rank d? for Σ of "approximately" rank d?

Run Demo_PCMIexe.m in Diffusion Geometry code for 1a and 1b.

2. Geometric Multi-Resolution Analysis

a. What is the approximation rate for GMRA for a smooth compact manifold of dimension d embedded in $\mathbb{R}^D$? How does it depend on the "curvature" of the manifold? How do d and D enter in the bounds? What happens for a union of manifolds?

b. Perform a back-of-the-envelope calculation for the sample complexity of empirical GMRA when the underlying distribution of the data is on a uniformly smooth manifold of dimension d. Assume that $\rho(C_{j,k}) \asymp 2^{-jd}$.

   i. assume that the tree construction works well up to a scale j such that every ball of radius 2^{-j} contains at least $O(1)$ points. If so, what is the maximal scale for which the tree construction can be performed reliably?

   ii. assume that the calculation of the empirical approximate affine tangent plane in each cell $C_{j,k}$ is reliable provided that $C_{j,k}$ contains at least $d \log d$ points. What is the maximal scale for which the approximate tangent plane may be estimated reliably?

   iii. on $C_{j,k}$ the least square approximation leads to a least squared error of order 2^{-4j}: perform bias/variance tradeoff to obtain a guess for the rate of approximation to be expected from empirical GMRA.

c. What are the obstacles to generalizing the above to the case when the manifold is not uniformly smooth?

d. Come up with examples of sets and measures on them which are not manifolds, and yet you would expect GMRA to perform well? Same question, but with GMRA to perform poorly?

Session III

1. Diffusion Wavelets

a. Give an example of a function on a circle graph that would be well-approximate by a short linear combination of wavelet coefficients, but not by a short linear combination of eigenfunctions of the Laplacian. Provide examples on more complicate graphs, and on graphs obtained as geometric graphs from samples from a manifold.

b. What are diffusion wavelets trying to accomplish, and why are these objectives partially incompatible with each other, and therefor hard to attain simultaneously?

c. How would you use diffusion wavelets to perform regression or semisupervised learning on graphs?

References

[1] 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)**, pp.7426-7431, 2005.

[2] A. V. Little, M. Maggioni and L. Rosasco, "Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature", *Applied and Computational Harmonic Analysis*, 2016.

[3] W. K. Allard, G. Chen and M. Maggioni, "Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis", *Applied and Computational Harmonic Analysis* **32(3)**, pp.435-462, 2012.

[4] M. Maggioni, S. Minsker and N. Strawn, "Multiscale dictionary learning: non-asymptotic bounds and robustness", to appear in *Journal of Machine Learning Research.*

[5] R. R. Coifman, and M. Maggioni, "Diffusion wavelets," *Applied and Computational Harmonic Analysis* **21(1)**, pp.53-94, 2006.