Geometry of Data and Biology

Mauro Maggioni

Introduction

The analysis of large high-dimensional data sets is a necessity in a wide variety of fields, including statistics, engineering and signal processing, physics, biology and medicine. While in the field of statistics data has always been at the center of attention, in the past several years the nature of many data sets has changed in a way that requires novel approaches, both from a theoretical and a practical perspective. Modern data sets may be very large but are very often high-dimensional, meaning each data point has a long list of attributes or coordinates, and this is often the case: This happens frequently in biological data (e.g., a genetic profile has easily more than 10^4 entries). While a large number n of data points is beneficial for statistical analysis, the high-dimension D of the data is a "curse" in the sense that, without further assumptions or model on the data, for many classical statistical inference and function approximation techniques to work n is required to scale exponentially in D, a truly gargantuan requirement (think about what 2^{10^4} looks like) [11]. Various types of assumptions on the data are usually made to avoid this curse, including parametric and nonparametric statistical models, as well as geometric models. These are not disjoint approaches but rather different languages to express modeling assumptions and provide a priori information about the structure of data. These hypotheses may often be interpreted geometrically in terms of imposing that the data is intrinsically low dimensional, and this property is to be exploited in learning algorithms and statistical tasks so that their performance depends on the intrinsic dimension d but not on the ambient dimension D.

Mauro Maggioni is professor of mathematics at Duke University. His email address is mauro@math.duke.edu

DOI: http://dx.doi.org/10.1090/noti1289

Graphs, Geometry, and Spectral Methods

We discuss here some aspects of one of these geometric frameworks that arose within the machine-learning community under the name of manifold learning. The basic idea is to capture the intrinsic geometry of data by constructing a graph connecting nearby data points and then study the geometry of such a graph.

Given data $X_n := \{x_i\}_{i=1}^n \subset \mathbb{R}^D$ one constructs a weighted graph G with vertex set X_n and edges on each pair (x_i, x_j) of weight

$$W_{ij} = e^{-\frac{||x_i - x_j||^2}{\sigma}} \tag{1}$$

(self-edges included) or other similarity measure (more on this below). Let D be the degree matrix, i.e., the diagonal matrix with $D_{x_i x_i} = \sum_{x_j} W_{x_i x_j}$. We consider the random walk associated to the Markov matrix $P = D^{-1}W$ and the normalized Laplacian $\mathcal{L} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. These objects are widely studied, in particular in the field of spectral graph theory [3], where the properties of eigenvalues and eigenvectors of P and $\mathcal{L}$ (they are closely related) are used to glean information about the geometry of the graph.

In *diffusion geometry* [7], [8], [5], [6] one focuses on the random walk P and studies relationships to both processes on graphs and to high-dimensional dynamical systems [4], [18], [22], both of which appear often in biology. Let $P\varphi_i = \lambda_i \varphi_i$ be the spectral decomposition of P, with eigenvalues sorted in decreasing order: $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq \cdots$. We consider the *diffusion map* $\Phi_{t,m}$, an embedding of G into $\mathbb{R}^m$, for some $t \geq 0$

and $m \geq 1$, defined by $\Phi_{t,m}(x) = (\sqrt{\lambda_i^t}\varphi_i(x))_{i=1}^m$. Various versions of this map have a long history in computer science, where they were used to embed graphs in the plane, and in differential geometry [2]. Its properties in general are still only partially understood, and for certain families of graphs this map may not have desirable properties (e.g., for expanders).

Consider the case when the points X_n are sampled independently from the normalized volume measure on a d-dimensional Riemannian manifold $\mathcal{M}$ in $\mathbb{R}^D$. The graph G and the operators constructed on it, such as P and $\mathcal{L}$, are then random with the samples. One may prove that the (random linear) operator $\mathcal{L}$ converges to the intrinsic Laplace-Beltrami operator on $\mathcal{M}$ as n tends to infinity.[1]

Spectral Clustering and Diffusion

In the case of graphs, the diffusion map $\Phi_{t,m}$ may be shown to have good metric properties when a suitable metric is defined on the graph. We define the diffusion distance on G at time t by

$$d_t(x,y) = ||P^t(x,\cdot) - P^t(y,\cdot)||_{L^2(G)},$$

where $||f||_{L^2(G)}^2 = \sum_{x\in G} |f(x)|^2$. In other words, the diffusion distance $d_t(x,y)$ between two points x, y is large if random walks of length t starting from x and y rarely reach the same point with similar probabilities.

It is easy to see, by spectral expansion, that

$$d_t(x,y)^2 = \sum_{i=1}^{n} \lambda_i^t |\varphi_i(x) - \varphi_i(y)|^2$$
$$= ||\Phi_{t,n}(x) - \Phi_{t,n}(y)||_{\mathbb{R}^n}^2 \cong ||\Phi_{t,m}(x) - \Phi_{t,m}(y)||_{\mathbb{R}^m}^2.$$

In other words, Euclidean distances in the range of $\Phi_{t,m}$ are comparable to diffusion distances on G. When is diffusion distance particularly useful? If the graph contains clusters C_1, C_2, one would expect that if the definition of cluster is "reasonable," few random walks of length t that started from an $x \in C_1$ would transition to a $y \in C_2$ (unless t is very large) and vice versa; thus the diffusion distance $d_t(x,y)$ ought to be large. This is in fact the case, for example, when conductance is used to define clusters: let

$$c(S) = \frac{\sum_{x\in S, y\notin S} W_{xy}}{\min\{\sum_{x,y\in S} W_{xy}, \sum_{x,y\notin S} W_{xy}\}}$$

be the conductance of $S \subseteq G$. A cluster may be defined as a set S with small conductance: if $c(S)$ is small, then the sum of the weights of edges connecting S to its complement is small compared to both the sum of the weights of edges within S and the sum of those within S^c. This implies that a random walk that started in S is trapped in S with high probability for a long time. More precisely, with probability at least $1/2$ a walk that started in S at random[2] stays in S up to time $O(1/c(S))$. So if the two clusters C_1 and C_2 have low conductance, then the diffusion distance between any $x \in C_1$ and $y \in C_2$ is large up to time $O(1/\max\{c(C_1), c(C_2)\})$. This implies that $\Phi_{t,m}$ will map C_1 and C_2 far from each other in Euclidean space, where they will have a chance of being separated by a simple boundary, e.g., a hyperplane. One of the most powerful and widely used clustering algorithms is *spectral clustering*, which dates to work in the 1970s [12], [13] (see also the overviews [19], [21]) and may be motivated by these ideas.

Connections to Biology

High-dimensional, large data sets collected in the biological sciences require, as discussed above, novel mathematical and statistical tools. We mention two examples, among many, where geometric models are particularly relevant.

First of all, *clustering* is a very common task in the analysis of biological data (see [14], [20], [15], [1] and references therein, among many, many others). A similarity function between the objects to be clustered (genes, proteins, genomes, philogenetic trees, etc.) is defined, and clustering or hierarchical clustering techniques with respect to these similarity measures are used to group the objects of interest (e.g., gene responses across patients) in order to discover structure among them. It is often difficult to have a priori information about the shape of the clusters, and tools such as spectral clustering that can accommodate the biologically defined similarity measures into a graph and do not require assumptions on the shapes of the clusters are an ideal tool.

Another, perhaps surprising, connection is with *molecular dynamics* (MD), often used to explore properties of biomolecules: the motion of a molecule in solution is often modeled by large systems of stochastic differential equations, with the stochasticity representing the effect of the solvent on the molecule. Leaving aside the (noneligible) shortcomings of such models, one obtains a first-order Langevin equation representing diffusion in a potential well:

$$dY_t = -\nabla U(Y_t)dt + \Sigma dB_t,$$

where Y_t is a $3N$-dimensional vector containing the three-dimensional coordinates of the N atoms in the molecule, U is a potential defined on the state space $\mathbb{R}^{3N}$, B_t is a standard Wigner process, and Σ determines the diffusion tensor. N is typically very large.

[1] *In a suitable sense, with high probability achieved, by choosing the σ in the construction of W in an appropriate fashion as a function of n and the intrinsic dimension d of $\mathcal{M}$ and up to an appropriate normalization* [7], [10].

[2] *More precisely, according to the stationary distribution of P.*

Often one is mostly interested in particular features of these systems, typically events occurring at large time scales, such as relative energy levels or transition rates between metastable states, which are local minima of the energy corresponding to regions of state space where the molecule spends a significant amount of time before hopping to another such state. In this situation it is natural to conjecture that there may be a small number of effective degrees of freedom of the molecule that determine its behavior at such large time scales. Geometrically speaking, there may be an intrinsically low-dimensional set around which large time trajectories accumulate; metastable states would correspond to "clusters" which are well separated in terms of (expected) time to transition among them.

This way of thinking suggests connections with the ideas of diffusion distance, clusters, and low-conductance sets described above. Several recent papers have made significant contributions to this line of thinking from a geometric perspective: first it was established [7], [8], [4] that, given a set of independent samples from long molecular dynamics trajectories, one may slightly modify the construction of the graph above to obtain a graph Laplacian that approximates not a manifold Laplacian but the Laplacian associated to the Fokker-Planck equation corresponding to the Langevin diffusion above, describing the evolution of the probability distribution of the position of the molecule. It is then natural to use diffusion maps to reduce the dimensionality of the state space to $\mathbb{R}^m$, $m \ll 3N$, and ask whether one may write or learn from data a Langevin equation in $\mathbb{R}^m$ with trajectories consistent with the original ones (see Figure 1). This type of model reduction is sought in order to glean information about these high-dimensional systems, but the crucial choice of low-dimensional embedding is usually done by hand by physical chemists using their expertise.

While promising, the standard diffusion map technique turned out to be rather sensitive to the choices of parameters (in particular, the parameter σ in the construction of the weights in the graph (1)). It was also not clear if the assumption that large-time trajectories did indeed concentrate around low-dimensional sets was a good model for real data. A novel technique called *multiscale SVD* [17] was developed to estimate in a robust way the intrinsic dimension of point clouds. The basic idea is to perform the SVD decomposition of data in a ball $B_z(r)$ centered at a data point z and having radius r and use the behavior of the singular values for fixed z and varying r to determine the intrinsic dimension of data and a suitable scale (value of r) at which the data around z is low dimensional. These constructions are inspired by classical work in geometric measure theory [16]. This construction was generalized to certain

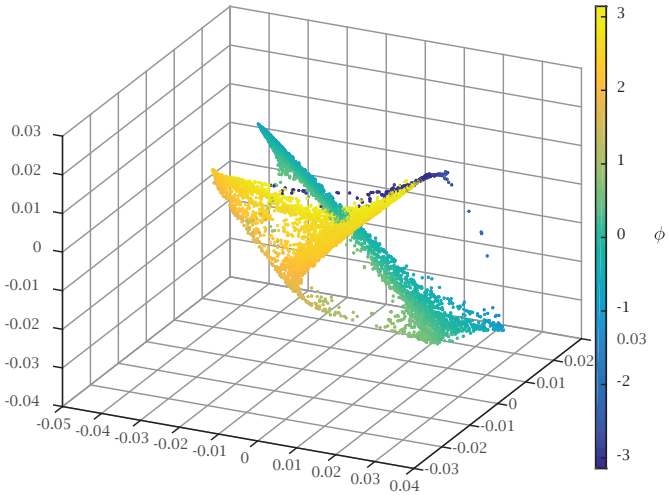

Figure 1. Diffusion map of configurations of a small peptide (alanine dipeptide) with $N = 12$ atoms, simulated with implicit solvent (i.e., instead of simulating all the water molecules, their "net force" is simulated by introducing noise and modifying the force field), computed with the algorithms described in [18], which take into account local dimensionality and the local scale where data is nearly "linear." Dense regions correspond to metastable states, thin connections to transition paths between those. The apparent self-intersection is an artifact of embedding from $D = 36$ to $m = 3$ dimensions and disappears when $m = 4$. The points are colored by one of the dihedral angles of the molecule, ϕ, known to be particularly important for the effective dynamic of the molecule. It is clearly a function of the first two diffusion coordinates constructed by the algorithm, and so is the second dihedral angle (not shown). The two most important metastable regions are revealed and well separated in the first diffusion coordinate; each of them has two subregions, which are separated by higher-order diffusion coordinates. As detailed in [18], a one-dimensional Langevin process may be estimated in the first diffusion coordinate that accurately reproduces key features of the large-time dynamics of the original system.

non-Euclidean spaces (such as the ones modeling the state space in molecular dynamics, which is naturally $\mathbb{R}^{3N}$ modulo the Galileo group) and deployed to unveil that indeed it is often the case, especially near crucial transition paths between metastable states, that the intrinsic dimension of long trajectories is indeed much smaller than the dimension of the state space [18]. Moreover, multiscale SVD yields an estimate of a "good local scale" $r^*(z)$ around each point z, which is roughly the largest r such that the data in $B_z(r)$ is both intrinsically low dimensional and well

approximated by a low-dimensional hyperplane (i.e., it is roughly flat). It turns out that, for molecular dynamics data, both the intrinsic dimension and this "good local scale" are often highly variable across regions of state space. It is natural to use this "good local scale" as the scale parameter σ (now a function of x_i instead of a constant) in the graph construction (1), yielding a robust generalization of diffusion maps called locally scaled diffusion maps [18], which has been used successfully to perform model reduction for several molecules of interest, yielding extremely low-dimensional and useful representations of large molecules, even in cases where it had been suggested that no such representations were possible [22].

Formidable challenges, both practical and theoretical, still exist. For example, these techniques require samples from long trajectories, which may be prohibitively expensive. Novel algorithms are being developed [9] to accelerate the simulation of molecular dynamics paths based on the dimension estimation and reduction above, given only a large number of short paths, whose calculation, while still expensive, is trivially parallelizable.

Conclusion

Biological data sets are large, high-dimensional, extremely diverse, and complex: novel mathematical, algorithmic, and interactive tools are needed to extract information, help visualize, and interpret the data collected. These mathematical constructions and algorithms are to data what instruments collecting data are to the physical world: they help capture, quantify, and inspire the discovery of the rules that govern their inputs.

Acknowledgements

The author would like to thank M. Reed for suggesting that the author write a section within Reed's article, which eventually took the form of this note. The author also thanks the referees, who made many suggestions that significantly improved the presentation. The author is grateful to AFOSR, ONR, and NSF, who supported much of the research work, cited below, by the author and his collaborators.

References

[1] N. Altemose, K. H. Miga, M. Maggioni, and H. F. Willard, Genomic characterization of large heterochromatic gaps in the human genome assembly, *PLoS Comput. Biol.* **10** (2014), e1003628.

[2] P. Bérard, G. Besson, and S. Gallot, Embedding Riemannian manifolds by their heat kernel, *Geom. Funct. Anal.* **4** (1994), 374–398.

[3] F. R. K. Chung, *Spectral Graph Theory*, vol. 92 of CBMS Regional Conference Series in Mathematics, published for the Conference Board of the Mathematical Sciences, Washington, DC, by the Amer. Math. Soc., Providence, RI, 1997.

[4] R. Coifman, I. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Diffusion maps, reduction coordinates and low dimensional representation of stochastic systems, *SIAM J.M.M.S.* **7** (2008), 842–864.

[5] R. Coifman and S. Lafon, Diffusion maps, *Appl. Comp. Harm. Anal.* **21** (2006), 5–30.

[6] R. Coifman and M. Maggioni, Diffusion wavelets, *Appl. Comp. Harm. Anal.* **21** (2006), 53–94 (Tech. Report YALE/DCS/TR-1303, Yale Univ., September 2004).

[7] 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, *Proc. Natl. Acad. Sci., USA* **102** (2005), 7426–7431.

[8] ———, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, *Proc. Natl. Acad. Sci., USA* **102** (2005), 7432–7438.

[9] M. Crosskey and M. Maggioni, Atlas: A geometric approach to learning high-dimensional stochastic systems near manifolds, 2014, submitted, arXiv: 1404.0667.

[10] M. J. Daniel Ting, Ling Huang, and Michael I. Jordan, An analysis of the convergence of graph Laplacians, arXiv: 1101.5435, 2011.

[11] D. Donoho, High-dimensional data analysis: The curses and blessings of dimensionality, *"Math Challenges of the 21st Century"*, UCLA, August 7–12, 2000, organized by the Amer. Math. Soc.

[12] W. E. Donath and A. J. Hoffman, Lower bounds for the partitioning of graphs, *IBM J. Res. Develop.* **17** (1973), 420–425.

[13] M. Fiedler, Algebraic connectivity of graphs, *Czechoslovak Math. J.* **23** (1973), 298–305.

[14] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, A survey of kernel and spectral methods for clustering, *Pattern Recognition* **41** (2008), 176–190.

[15] D. J. Higham, G. Kalna, and M. Kibble, Spectral clustering and its use in bioinformatics, *J. Comput. Appl. Math.* **204** (2007), 25–37. Special issue dedicated to Professor Shinnosuke Oharu on the occasion of his 65th birthday.

[16] P. W. Jones, Rectifiable sets and the traveling salesman problem, *Invent. Math.* **102** (1990), 1–15.

[17] A. V. Little, M. Maggioni and L. Rosasco, *Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature*, tech. report, MIT-CSAIL-TR-2012-029/CBCL-310, MIT, Cambridge, MA, September 2012.

[18] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, Determination of reaction coordinates via locally scaled diffusion map, *J. Chem. Phys.* (2011), 124116.

[19] D. Spielman and S. Teng, Spectral partitioning works: Planar graphs and finite element meshes, *FOCS* (1996, 96–105.

[20] D. Tritchler, S. Fallah, and J. Beyene, A spectral clustering method for microarray data, *Comp. Stat. Data Anal.* **49** (2005), 63–76.

[21] U. von Luxburg, A tutorial on spectral clustering, *CoRR*, abs/0711.0189 (2007).

[22] W. Zheng, M. A. Rohrdanz, M. Maggioni, and C. Clementi, Polymer reversal rate calculated via locally scaled diffusion map, *J. Chem. Phys.* (2011), 144108.