Diffusion maps

By Ronald R. Coifman, Stéphane Lafon

Weiqiang Jing April 17, 2019
Diffusion map: A framework based upon diffusion processes for finding meaningful geometric descriptions of data sets.

- Dimensionality reduction
- Goal: change the representation of data sets, originally in a form involving a large number of variables, into a low-dimensional description using only a small number of free parameters.

A great deal of attention has been recently paid to the so-called “kernel eigenmap methods” such as local linear embedding, Laplacian eigenmaps, Hessian eigenmaps. The remarkable idea emerging from these papers is that eigenvectors of Markov matrices can be thought of as coordinates on the data set.
Two major advantages over classical dimensionality reduction methods (PCA, classical multidimensional scaling):

- They are nonlinear
- They preserve local structures

In this paper, they show that all these kernel eigenmap methods constitute special cases of a general framework based on diffusion processes.

They use the eigenfunctions of a Markov matrix defining a random walk on the data to obtain new descriptions of data sets via a family of mappings that they term “diffusion maps.”
Given $X$ as the data set, suppose given a "kernel" $k : X \times X \rightarrow \mathbb{R}$ that satisfies:

1. $k$ is symmetric: $k(x, y) = k(y, x)$
2. $k$ is positivity preserving: $k(x, y) \geq 0$
Given graph defined by \((X, k)\), one can construct a reversible Markov chain on \(X\). This is known as normalized graph Laplacian construction.

Local measure of degree \(d(x) = \int_X k(x, y)d\mu(y)\)

Define: \(p(x, y) = \frac{k(x, y)}{d(x)}\)

\[\int_X p(x, y)d\mu(y) = 1\] means it can be viewed as the transition kernel of a Markov chain on \(X\)

\[\text{Pf}(x) = \int_X a(x, y)f(y)d\mu(y)\]
Powers of $P$ and multiscale geometric analysis of $X$
Spectral analysis of the Markov chain

A classical way to describe the powers of an operator is to employ the language of spectral theory, namely eigenvectors and eigenvalues.

The random walk that we have constructed exhibits very particular mathematical properties:

- The Markov chain has a stationary distribution given by

\[ \pi(y) = \frac{d(y)}{\sum_{z \in X} d(z)} \]

- The chain is reversible

\[ \pi(x)p(x, y) = \pi(y)p(y, x) \]

- If \( X \) is finite and the graph of the data is connected, then the chain is ergodic.
Diffusion distances and diffusion maps

introduce the family of diffusion distances $\{D_t\}_{t \in \mathbb{N}}$ given by

$$D_t(x, y)^2 \triangleq \|p_t(x, \cdot) - p_t(y, \cdot)\|_{L^2(X, d\mu/\pi)}^2 = \int_X (p_t(x, u) - p_t(y, u))^2 \frac{d\mu(u)}{\pi(u)}$$

$D_t(x, y)$ is a functional weighted $L^2$ distance between the two posterior distributions $u \mapsto p_t(x, u)$ and $u \mapsto p_t(y, u)$. $D_t(x, y)$ is small if there is a large number of short paths connecting $x$ and $y$ i.e if there is a large probability of transition from $x$ to $y$ and vice versa.
There are three main interesting features of the diffusion distance:

- This distance emphasizes the notion of a cluster.
- This number is very robust to noise perturbation, unlike the geodesic distance.
- This distance takes into account all evidences relating $x$ and $y$.

It is shown that $D_t(x, y)$ can be computed using the eigenvectors and eigenvalues of $P$:

$$D_t(x, y) = \left( \sum_{l \geq 1} \lambda_l^{2t} (\psi_l(x) - \psi_l(y))^2 \right)^{\frac{1}{2}}$$
Diffusion distances and diffusion maps

Compute finite number of terms to a preset accuracy \( \delta > 0 \), if we define:

\[
s(\delta, t) = \max \left\{ l \in \mathbb{N} \text{ such that } |\lambda_l|^t > \delta |\lambda_1|^t \right\}
\]

then, up to relative precision \( \delta \), we have

\[
D_t(x, y) = \left( \sum_{l=1}^{s(\delta, t)} \lambda_l^{2t} (\psi_l(x) - \psi_l(y))^2 \right)^{1/2}
\]

Therefore the family of diffusion maps \( \{\Psi_t\}_{t \in \mathbb{N}} \) given by:

\[
\Psi_t(x) \triangleq \begin{pmatrix}
\lambda_1^t \psi_1(x) \\
\lambda_2^t \psi_2(x) \\
\vdots \\
\lambda_{s(\delta, t)}^t \psi_{s(\delta, t)}(x)
\end{pmatrix}
\]

\( \Psi_t : X \rightarrow \mathbb{R}^{s(\delta, t)} \)
The connection between diffusion maps and diffusion distances can be summarized as follows:

**Proposition**

The diffusion map $\Psi_t$ embeds the data into the Euclidean space $\mathbb{R}^{s(\delta, t)}$ so that in this space, the Euclidean distance is equal to the diffusion distance (up to relative accuracy $\delta$), or equivalently,

$$\|\Psi_t(x) - \Psi_t(y)\| = D_t(x, y)$$
They generated a collection of images of the word “3D” viewed under different angles and given in no particular order. We formed a diffusion matrix $P$ based on a Gaussian-weighted graph and computed the diffusion coordinates.
In general, for a given time $t$, the number of eigenvectors used for parametrizing the data is equal to the number of eigenvalues to the powers of $t$ that have a magnitude greater than a given threshold $\delta$. Therefore, the dimensionality of the embedding depends on both $t$ and the decay of the spectrum of $P$.

Two extreme cases:

- All the nodes are disconnected $P$ is equal to the identity operator and thus to a flat spectrum.

- All nodes are connected $P$ has one eigenvalue equal to 1, and all other eigenvalues are equal to 0.

The decay of the spectrum is therefore a measure of the connectivity of points in the graph.
Anisotropic diffusions for points in $\mathbb{R}^n$

Question: What is the influence of the density of the points and of the geometry of the possible underlying data set over the eigenfunctions and spectrum of the diffusion?

To address this type of question, we now introduce a family of anisotropic diffusion processes that are all obtained as small-scale limits of a graph Laplacian jump process. This family is parameterized by a number $\alpha \in \mathbb{R}$ which can be tuned up to specify the amount of influence of the density in the infinitesimal transitions of the diffusion.
Anisotropic diffusions for points in \( \mathbb{R}^n \)

Fix the notation:
Let \( \mathcal{M} \) be a compact \( C^\infty \) submanifold of \( \mathbb{R}^n \). The heat diffusion on \( \mathcal{M} \) is the diffusion process whose infinitesimal generator is the Laplace-Beltrami operator \( \Delta \). Let the Neumann heat kernel be denoted \( e^{-t\Delta} \). The operator \( \Delta \) has eigenvalues and eigenfunctions on \( \mathcal{M} \):

\[
\Delta \phi_l = \sqrt{\lambda_l} \phi_l
\]

Let:

\[
E_K = \text{Span} \{ \phi_l, 0 \leq l \leq K \}
\]

Another expression for the Neumann heat kernel is given by

\[
e^{-t\Delta} = \lim_{s \to +\infty} \left( I - \frac{\Delta}{s} \right)^{st} = \sum_{l \geq 0} e^{-t\sqrt{\lambda_l}} \phi_l(x) \phi_l(y)
\]

Assume that the data set \( X \) is the entire manifold, and let \( q(x) \) be the density of the points on \( \mathcal{M} \).
Construction of a family of diffusions

(1) Fix $\alpha \in \mathbb{R}$ and a rotation-invariant kernel $k_\varepsilon(x, y) = h\left(\frac{\|x-y\|^2}{\varepsilon}\right)$

(2) Let:

$$q_\varepsilon(x) = \int_X k_\varepsilon(x, y)q(y)dy$$

and form the new kernel

$$k_\varepsilon^{(\alpha)}(x, y) = \frac{k_\varepsilon(x, y)}{q_\varepsilon^\alpha(x)q_\varepsilon^\alpha(y)}$$

(3) Apply the weighted graph Laplacian normalization to this kernel by setting

$$d_\varepsilon^{(\alpha)}(x) = \int_X k_\varepsilon^{(\alpha)}(x, y)q(y)dy$$

and by defining the anisotropic transition kernel

$$p_\varepsilon,\alpha(x, y) = \frac{k_\varepsilon^{(\alpha)}(x, y)}{d_\varepsilon^{(\alpha)}(x)}$$

Let $P_\varepsilon,\alpha f(x) = \int_X p_\varepsilon,\alpha(x, y)f(y)q(y)dy$
Construction of a family of diffusions

**Theorem**

Let

\[ L_{\varepsilon, \alpha} = \frac{I - P_{\varepsilon, \alpha}}{\varepsilon} \]

be the infinitesimal generator of the Markov chain. Then for a fixed \( K > 0 \), we have on \( E_K \)

\[ \lim_{\varepsilon \to 0} L_{\varepsilon, \alpha} f = \Delta \left( f q^{1-\alpha} \right) - \Delta \left( q^{1-\alpha} \right) \frac{f}{q^{1-\alpha}} \]

In other words, the eigenfunctions of \( P_{\varepsilon, \alpha} \) can be usproximate those of the following symmetric Schrödinger operator:

\[ \Delta \phi - \Delta \left( q^{1-\alpha} \right) \frac{\phi}{q^{1-\alpha}} \]

where \( \phi = f q^{1-\alpha} \).
Three cases

The case $\alpha = 0$ : normalized graph Laplacian on isotropic weights From the previous theorem, the corresponding infinitesimal operator is given by

$$\Delta \phi - \frac{\Delta q}{q} \phi$$

The case $\alpha = \frac{1}{2}$ : Fokker-Planck diffusion When $\alpha = \frac{1}{2}$, the asymptotic infinitesimal generator reduces to

$$\Delta \phi - \frac{\Delta (\sqrt{q})}{\sqrt{q}} \phi$$

Let us write $q = e^{-U}$, then the generator becomes

$$\Delta \phi - \left( \frac{\|\nabla U\|^2}{4} - \frac{\Delta U}{2} \right) \phi$$

It can be shown that a simple conjugation of this specific Schrödinger operator leads to the forward Fokker–Plank equation

$$\frac{\partial q}{\partial t} = \nabla \cdot (\nabla q + q \nabla U)$$
One example

The case $\alpha = 1$: approximation of the heat kernel

**Proposition**

We have:
\[
\lim_{\varepsilon \to 0} L_{\varepsilon,1} = \Delta
\]

Furthermore, for any $t > 0$, the Neumann heat kernel $e^{-t\Delta}$ can be approximated on $L^2(M)$ by $P_{\varepsilon,1}^{t/\varepsilon}$

\[
\lim_{\varepsilon \to 0} P_{\varepsilon,1}^{t/\varepsilon} = e^{-t\Delta}
\]

By setting $\alpha = 1$, the infinitesimal generator is simply the Laplace-Beltrami operator $\Delta$. 
Parametrization of data and dimensionality reduction

Weiqiang Jing
Diffusion maps
April 17, 2019
This paper introduced a set of tools, the diffusion maps and distances, and explained their construction from Markov processes defined on a set $X$. The diffusion coordinates provide representation of the local geometry of the graph of the data. These coordinates allow us to parametrize the data set, and they also reflect the connectivity within the graph. Dimensionality reduction is achieved thanks to the decay of the eigenvalues of the diffusion operator.

In the case of data points in the Euclidean space $\mathbb{R}^n$, they constructed a one-parameter family of diffusion kernels that capture different features of the data. In particular, we showed how to design a kernel that reproduces the diffusion induced by a Fokker–Planck equation. This allows to compute the long-time dynamics of some stochastic differential system.