Diffusion Maps

University of Vienna

Peter Mühlbacher

October 1, 2014

Abstract

Contents

Introduction

The contentual layout follows that of Coifman and Lafon (2006) and Belkin (2003) for a concrete implementation. Some definitions, theorems and examples will be taken from these works as well.

0.1 Dimensionality Reduction

0.1.1 The Problem

0.1.2 Existing Approaches

Linear PCA

Kernel PCA

Chapter 1

Diffusion Maps

1.1 Diffusion Kernels

1.1.1 Motivation

In this section it will be elaborated on why global structures need not be preserved and how this leads to diffusion processes.

Preservation of Global Structures

Suppose data points $x_i \in \mathbb{R}^k$ are generated by a low dimensional parameter $\theta_i \in \mathbb{R}^{k'}$, $k' \ll k$ given a map $\Phi : \mathbb{R}^{k'} \to \mathbb{R}^k$. One problem is the (numerical) smoothness of Φ which is not necessarily given. Another problem is that for large k the euclidian distance is no longer a meaningful measure as the volume of the k-dimensional unit ball $\frac{\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2}+1)}$ converges to 0 as $lim_{k\to\infty}$ which is also known as the *curse of dimensionality*. One may conclude that large distances in the ambient space need not necessarily be preserved as they do not hold much information except that x_i is not "very close" to x_j.

Analogous to Riemannian manifolds (where metric tensors, inducing an inner product on the tangent space and a metric via the exponential map, define the manifold's geometry) we focus solely on local distances in order to recover intrinsic global structures.

A Dual Approach

From inverse problems in spectral geometry (e.g. "Can One Hear the Shape of a Drum?") it is known that much[1] of the geometry of a given set Γ can be derived from the analysis of functions defined on Γ.

In this work eigenvalues and eigenfunctions of averaging operators, i.e., operators whose kernel corresponds to transition probabilities of a Markov process, will be studied in order to define a diffusion map which embeds the data into a Euclidian space where the Euclidian distance is just the diffusion metric.

1.1.2 Construction of a Random Walk on the Data

Definitions

Let $(\Gamma, \mathcal{A}, \mu)$ be a measure space. In practical applications Γ is the given data set consisting of finitely many data points and μ is the counting measure to represent the distribution of the points in the data set. In addition, suppose we are given a symmetric kernel $k : \Gamma \times \Gamma \to \mathbb{R}^+$ which defines the local geometry of Γ.

Examples

Usually Γ is either a subset of the Euclidian space or a weighted graph.

In the first case it seems natural to write k as a function of the Euclidian distance $\nu(||x-y||)$.

In the second case let $b(x,y)$ be the associated adjacancy matrix, that is, $b(x,y)=1$ if there is an edge going from x to y, and $b(x,y)=0$ otherwise. The kernel b defines a notion of neighborhood for each point, and also a non-symmetric distance given by $1-b(x,y)$. Clearly b is not symmetric in general, but we can consider

$$k_1(x,y) := \int_\Gamma b(x,u)b(y,u)d\mu(u)$$

$$k_2(x,y) := \int_\Gamma b(u,x)b(u,y)d\mu(u)$$

where $k_1(x,y)$ counts the number of common neighbors to x and y, while $k_2(x,y)$ counts the number of nodes for which x and y are common neighbors.

[1] the most famous example being Weyl's proof of $\#\{\lambda_k : \lambda_k < \lambda\} \sim \frac{\text{area}(\Gamma)}{2\pi}\lambda$ as $\lambda \to \infty$. (Canzani (2013))

Normalized Graph Laplacian Construction

Generally, such a kernel represents some notion of affinity between points of Γ and thus one can think of the data points as being the nodes of a symmetric graph whose weight function is specified by k. From the graph defined by (Γ, k), one can construct a reversible Markov chain on Γ.

To normalize the kernel, define

Definition 1.1.

$$v^2(x) = \int_\Gamma k(x,y) d\mu(y)$$

and

Definition 1.2.

$$p(x,y) = \frac{k(x,y)}{v^2(x)}.$$

$p(x,y)$ is no longer symmetric, but inherited the positivity and now satisfies a conservation property:

$$\int_\Gamma p(x,y) d\mu(y) = 1$$

As a result the matrix $P := (p(i,j))_{i,j}$ is stochastic and can be interpreted as the transition matrix of a homogeneous Markov process on Γ. In spectral graph theory $\mathbb{I} - P$ is commonly referred to as normalized, weighted graph Laplacian. This naming will be justified with the following theorem:

Theorem 1.1.1 (Convergence of Graph Laplacian)**.**

Proof. □

To investigate the spectral properties of the corresponding integral operator P defined by $Pf(x) = \int_\Gamma p(x,y)f(y)d\mu(y)$ it is beneficial to examine the symmetric, conjugated Operator A.

First, notice that by setting $a(x,y) = \frac{k(x,y)}{v(x)v(y)} = v(x)p(x,y)\frac{1}{v(y)}$ one obtains a symmetric form and thus a symmetric operator $Af(x) = \int_\Gamma a(x,y)f(y)d\mu(y)$ which will be referred to as diffusion operator from now on.

1.1.3 Diffusion Kernels

Theorem 1.1.2 (Spectral Properties of the Diffusion Operator)**.** *The diffusion operator A with kernel a is bounded from $L^2(\Gamma, d\mu)$ into itself, symmetric and positive semi-definite.*

Moreover, its norm is

$$||A|| = 1$$

and is taken by the eigenfunction

$$Av = v$$

.

Proof. Let $f \in L^2(\Gamma, d\mu)$. We have:

$$\langle Af, f\rangle = \int_{\Gamma^2} k(x,y)\frac{f(x)}{v(x)}\frac{f(y)}{v(y)}d\mu(x)d\mu(y) \tag{1.1}$$

.

Applying the Cauchy-Schwartz inequality we get:

$$\begin{aligned}\Big|\textstyle\int_\Gamma k(x,y)\frac{f(y)}{v(y)}d\mu(y)\Big| &= \Big(\textstyle\int_\Gamma k(x,y)d\mu(y)\Big)^{\frac{1}{2}}\Big(\textstyle\int_\Gamma k(x,y)\frac{f(y)^2}{v(y)^2}d\mu(y)\Big)^{\frac{1}{2}} \\ &= v(x)(\textstyle\int_\Gamma k(x,y)\frac{f(y)^2}{v(y)^2}d\mu(y))^{\frac{1}{2}}\end{aligned}$$

Hence:

$$\langle Af, f\rangle \leq \int_\Gamma |f(x)|\Big(\int_\Gamma k(x,y)\frac{f(y)^2}{v(y)^2}d\mu(y)\Big)^{\frac{1}{2}}d\mu(x)$$

and by using the Cauchy-Schwartz inequality once again:

$$\langle Af, f\rangle \leq ||f||\Big(\int_{\Gamma^2}\frac{k(x,y)}{v(y)^2}f(y)^2d\mu(y)d\mu(x)\Big)^{\frac{1}{2}} = ||f||^2$$

by symmetry of the kernel which, in combination with 1.1, also implies the positivity of A.

Plugging in v for f it follows immediately that the eigenvalue 1 is actually obtained and v is an eigenfunction. □

Theorem 1.1.3 (Spectral Decomposition of the Diffusion Kernel)**.** *Assuming A is compact*[2] *and $A\phi_l = \lambda_l\phi_l$ we may write the kernel as*

try to generalize lemma 3.4 of **?**

$$a(x,y) = \sum_{l\geq 0}\lambda_l\phi_l(x)\phi_l(y)$$

with $\lambda_0 = 1$ and $\lim_{l\to\infty}\lambda_l = 0$ monotonically.

[2] which is no constraint in practice since data is finite

Proof. First, note that A being compact implies that the spectrum is discrete and the sum thus is well defined. By A being symmetric and compact the spectral theorem applies and we get that there exists a sequence of real eigenvalues λ_l converging to 0. The corresponding normalized eigenvectors ϕ_l form an orthonormal set and every $f \in L^2(\Gamma, d\mu)$ can be written as

$$f = \sum_{l \geq 0} \langle \phi_l, f \rangle \phi_l + h$$

where $h \in Ker(A)$.

It follows that

$$Af(x) = \int_\Gamma a(x,y) f(y) d\mu(y) = \sum_{l \geq 0} \lambda_l \int_\Gamma \phi_l(y) f(y) d\mu(y) \ \phi_l(x)$$

which, by linearity of the integral and comparison of components, is just what we were looking for. □

“Komponentenvergleich auf Englisch finden

From definition of $a(x,y)$ we see that

$$p(x,y) = \sum_{l \geq 0} \lambda_l \underbrace{\frac{\phi_l(x)}{v(x)}}_{=:\psi_l(x)} \phi_l(y) v(y) \tag{1.2}$$

which enables us to efficiently compute tth powers p_t of p.

There are two ways to interpret p_t:

1. p_t has a probabilistic interpretation as the probability for a Markov chain with transition matrix P to reach y from x in t steps.

2. the dual point of view is that of the functions defined on the data. The kernel p_t can be viewed as a bump or more precisely, if $x \in \Gamma$ is fixed, then $p_t(x, \cdot)$ is a bump function centered at x and of width increasing with t which intuitively captures the idea of diffusion.

1.1.4 Embedding in the Euclidian Space

Definition 1.3. *Let*

$$D_t(x,y)^2 = ||p_t(x, \cdot) - p_t(y, \cdot)||^2_{L^2(\Gamma, d\mu/v)} = \int_\Gamma \big(p_t(x,u) - p_t(y,u)\big)^2 \frac{d\mu(u)}{v(u)}$$

be the family of diffusion distances parameterized by t.

For a fixed value of t D_t defines a distance on the set Γ which is small only if there is a large number of small paths connecting x and y (i.e. if there is a large probability of getting from x to y in t steps). It thus emphasizes the notion of a cluster.

Another property following from the summation over all possible paths is that this distance is very robust to noise perturbation (in contrast to the geodesic distance).

Theorem 1.1.4 (A Numerically Feasible Representation)**.**

$$D_t(x,y) = \left(\sum_{l \geq 0} \left(\lambda_l^l (\psi_l(x) - \psi_l(y)) \right)^2 \right)^{\frac{1}{2}}$$

Proof. $\{\phi_l\}_{l\geq 0}$ forming an orthonormal basis for $L^2(\Gamma, d\mu)$ implies that $\{\phi_l v\}_{l\geq 0}$ is an orthonormal basis for $L^2(\Gamma, d\mu/v)$ and thus for x fixed 1.2 may be seen as orthogonal expansion of the function $y \mapsto p_t(x,y)$ into the basis $\{\phi_l v\}_{l\geq 0}$. The coefficients are given by $\{\lambda_l^t \psi_l(x)\}_{l\geq 0}$. The statement follows directly using the Pythagorean theorem. □

An imidiate consequence is that the diffusion distance is well approximable and that it converges towards a function of (numerical) rank 1 as $t \to \infty$ because of the vanishing influence of all eigenvectors with eigenvalues < 1.

One possible interpretation is that $D_t(x,y)$ measures the distance between bumps of "magnitude" t being centered around two points x and y. As t gets larger so does the size of the supports and the number of eigenfunctions needed to calculate $D_t(x,y)$ decreases. This number is related to the minimum number of bumps necessary to cover the set X (like in Weyl's asymptotic law for the decay of the spectrum).

In order to calculate $D_t(x,y)$ to a preset accuracy $\delta > 0$ with a finite number of terms we set

$$s_t(\delta) = \max\{l \in \mathbf{N} : \lambda_l^t \ > \delta \lambda_1^t\}$$

so that, up to relative precision δ

$$D_t(x,y) = \left(\sum_{l=0}^{s_t(\delta)} \left(\lambda_l^l (\psi_l(x) - \psi_l(y)) \right)^2 \right)^{\frac{1}{2}}. \tag{1.3}$$

Definition 1.4. *Let* $\{\Psi_t\}_{t\in\mathbf{N}}$,

$$\Psi_t(x) = \begin{pmatrix} \lambda_1^t\psi_1(x) \\ \lambda_2^t\psi_2(x) \\ \vdots \\ \lambda_{s_t(\delta)}^t\psi_{s_t(\delta)}(x) \end{pmatrix}$$

be the family of diffusion maps. Each component of $\Psi_t(x)$ *is termed diffusion coordinate.*

According to 1.3 diffusion maps embed data in a Euclidian space in such a way that the Euclidian distance equals the diffusion distance up to a relative error δ.

Appendix A

Special Cases

A.1 Laplacian Eigenmaps

Bibliography

Belkin, M. (2003), *Problems of Learning on Manifolds.* Ph.D. thesis, The University of Chicago.

Canzani, Y. (2013), “Analysis on manifolds via the laplacian.”, URL `http://www.math.harvard.edu/~canzani/math253/Laplacian.pdf`.

Coifman, Ronald R. and Stéphane Lafon (2006), “Diffusion maps.” *Applied and Computational Harmonic Analysis*, 21, 5 – 30, URL `http://www.sciencedirect.com/science/article/pii/S1063520306000546`. Special Issue: Diffusion Maps and Wavelets.