10

Diffusion Maps - a Probabilistic Interpretation for Spectral Embedding and Clustering Algorithms

Boaz Nadler[1], Stephane Lafon[2,3], Ronald Coifman[3], and Ioannis G. Kevrekidis[4]

[1] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel, boaz.nadler@weizmann.ac.il
[2] Google, Inc.
[3] Department of Mathematics, Yale University, New Haven, CT, 06520-8283, USA, coifman@math.yale.edu
[4] Department of Chemical Engineering and Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA, yannis@princeton.edu

Summary. Spectral embedding and spectral clustering are common methods for non-linear dimensionality reduction and clustering of complex high dimensional datasets. In this paper we provide a diffusion based probabilistic analysis of algorithms that use the normalized graph Laplacian. Given the pairwise adjacency matrix of all points in a dataset, we define a random walk on the graph of points and a *diffusion distance* between any two points. We show that the diffusion distance is equal to the Euclidean distance in the embedded space with all eigenvectors of the normalized graph Laplacian. This identity shows that *characteristic relaxation times and processes* of the random walk on the graph are the key concept that governs the properties of these spectral clustering and spectral embedding algorithms. Specifically, for spectral clustering to succeed, a necessary condition is that the mean exit times from each cluster need to be significantly larger than the largest (slowest) of all relaxation times inside all of the individual clusters. For complex, multiscale data, this condition may not hold and multiscale methods need to be developed to handle such situations.

10.1 Introduction

Clustering and low dimensional representation of high dimensional data sets are important problems in many diverse fields. In recent years various spectral methods to perform these tasks, based on the eigenvectors of adjacency matrices of graphs on the data have been developed, see for example [1]-[12] and references therein. In the simplest version, known as the normalized graph

Laplacian, given n data points $\{\boldsymbol{x}_i\}_{i=1}^n$ where each $\boldsymbol{x}_i \in \mathbb{R}^p$ (or some other normed vector space), we define a pairwise similarity matrix between points, for example using a Gaussian kernel with width σ^2,

$$W_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma^2}\right) , \tag{10.1}$$

and a diagonal normalization matrix $D_{ii} = \sum_j W_{ij}$. Many works propose to use the first few eigenvectors of the normalized eigenvalue problem $W\phi = \lambda D\phi$, or equivalently of the matrix

$$M = D^{-1}W , \tag{10.2}$$

either as a basis for the low dimensional representation of data or as good coordinates for clustering purposes. Although eq. (1) is based on a Gaussian kernel, other kernels are possible, and for actual datasets the choice of a kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ can be crucial to the method's success.

The use of the first few eigenvectors of M as good coordinates is typically justified with heuristic arguments or as a relaxation of a discrete clustering problem [3]. In [6, 7] Belkin and Niyogi showed that, when data is uniformly sampled from a low dimensional manifold of $\mathbb{R}^p$, the first few eigenvectors of M are discrete approximations of the eigenfunctions of the Laplace-Beltrami operator on the manifold, thus providing a mathematical justification for their use in this case. We remark that a compact embedding of a manifold into a Hilbert space via the eigenfunctions of the Laplace-Beltrami operator was suggested in differential geometry, and used to define distances between manifolds [13]. A different theoretical analysis of the eigenvectors of the matrix M, based on the fact that M is a stochastic matrix representing a random walk on the graph was described by Meilă and Shi [14], who considered the case of piecewise constant eigenvectors for specific lumpable matrix structures. Additional notable works that considered the random walk aspects of spectral clustering are [10, 15], where the authors suggest clustering based on the average commute time between points, [16, 17] which considered the relaxation process of this random walk, and [18, 19] which suggested random walk based agglomerative clustering algorithms.

In this paper we present a unified probabilistic framework for the analysis of spectral clustering and spectral embedding algorithms based on the normalized graph Laplacian. First, in section 10.2 we define a distance function between any two points based on the random walk on the graph, which we naturally denote the *diffusion distance*. The diffusion distance depends on a time parameter t, whereby different structures of the graph are revealed at different times. We then show that the non-linear embedding of the nodes of the graph onto the eigenvector coordinates of the normalized graph Laplacian, which we denote as the *diffusion map*, converts the diffusion distance between the nodes into Euclidean distance in the embedded space. This identity provides a probabilistic interpretation for such non-linear embedding algorithms.

It also provides the key concept that governs the properties of these methods, the characteristic relaxation times and processes of the random walk on a graph. Properties of spectral embedding and spectral clustering algorithms in light of these characteristic relaxation times are discussed in sections 10.3 and 10.4. We conclude with summary and discussion in section 10.5. The main results of this paper were first presented in [20] and [24].

10.2 Diffusion Distances and Diffusion Maps

The starting point of our analysis, as also noted in other works, is the observation that the matrix M is adjoint to a symmetric matrix

$$M_s = D^{1/2} M D^{-1/2} \,. \tag{10.3}$$

Thus, the two matrices M and M_s share the same eigenvalues. Moreover, since M_s is symmetric it is diagonalizable and has a set of n real eigenvalues $\{\lambda_j\}_{j=0}^{n-1}$ whose corresponding eigenvectors $\{\boldsymbol{v}_j\}$ form an orthonormal basis of $\mathbb{R}^n$. We sort the eigenvalues in decreasing order in absolute value, $|\lambda_0| \geq |\lambda_1| \geq \ldots \geq |\lambda_{n-1}|$. The left and right eigenvectors of M, denoted ϕ_j and ψ_j are related to those of M_s according to

$$\phi_j = \boldsymbol{v}_j D^{1/2}, \qquad \psi_j = \boldsymbol{v}_j D^{-1/2} \,. \tag{10.4}$$

Since the eigenvectors $\boldsymbol{v}_j$ are orthonormal under the standard dot product in $\mathbb{R}^n$, it follows that the vectors ϕ_i and ψ_j are bi-orthonormal

$$\langle \phi_i, \psi_j \rangle = \delta_{ij} \,, \tag{10.5}$$

where $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ is the standard dot product between two vectors in $\mathbb{R}^n$. We now utilize the fact that by construction M is a stochastic matrix with all row sums equal to one, and can thus be interpreted as defining a random walk on the graph. Under this view, M_{ij} denotes the transition probability from the point $\boldsymbol{x}_i$ to the point $\boldsymbol{x}_j$ in one time step,

$$\Pr\{\boldsymbol{x}(t+1) = \boldsymbol{x}_j \,|\, \boldsymbol{x}(t) = \boldsymbol{x}_i\} = M_{ij} \,. \tag{10.6}$$

We denote by $p(t, \boldsymbol{y}|\boldsymbol{x})$ the probability distribution of a random walk landing at location $\boldsymbol{y}$ at time t, given a starting location $\boldsymbol{x}$ at time $t = 0$. In terms of the matrix M, this transition probability is given by $p(t, \boldsymbol{y}|\boldsymbol{x}_i) = \boldsymbol{e}_i M^t$, where $\boldsymbol{e}_i$ is a row vector of zeros with a single entry equal to one at the i-th coordinate.

For ε large enough all points in the graph are connected, so that M is an irreducible and aperiodic Markov chain. It has a unique eigenvalue equal to 1, with the other eigenvalues strictly smaller than one in absolute value. Then, regardless of the initial starting point $\boldsymbol{x}$,

$$\lim_{t\to\infty} p(t, \boldsymbol{y}|\boldsymbol{x}) = \phi_0(\boldsymbol{y}) \ , \tag{10.7}$$

where ϕ_0 is the left eigenvector of M with eigenvalue $\lambda_0 = 1$, explicitly given by

$$\phi_0(\boldsymbol{x}_i) = \frac{D_{ii}}{\sum\limits_j D_{jj}} \ . \tag{10.8}$$

This eigenvector has a dual interpretation. The first is that ϕ_0 is the stationary probability distribution on the graph, while the second is that $\phi_0(\boldsymbol{x})$ is a density estimate at the point $\boldsymbol{x}$. Note that for a general shift invariant kernel $K(\boldsymbol{x} - \boldsymbol{y})$ and for the Gaussian kernel in particular, ϕ_0 is simply the well known Parzen window density estimator [21].

For any finite time t, we decompose the probability distribution in the eigenbasis $\{\phi_j\}$

$$p(t, \boldsymbol{y}|\boldsymbol{x}) = \phi_0(\boldsymbol{y}) + \sum_{j\geq 1} a_j(\boldsymbol{x})\lambda_j^t \phi_j(\boldsymbol{y}) \ , \tag{10.9}$$

where the coefficients a_j depend on the initial location $\boldsymbol{x}$. The bi-orthonormality condition (10.5) gives $a_j(\boldsymbol{x}) = \psi_j(\boldsymbol{x})$, with $a_0(\boldsymbol{x}) = \psi_0(\boldsymbol{x}) = 1$ already implicit in (10.9).

Given the definition of the random walk on the graph, it is only natural to quantify the similarity between any two points according to the evolution of probability distributions initialized as delta functions on these points. Specifically, we consider the following distance measure at time t,

$$\begin{aligned} D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) &= \|p(t, \boldsymbol{y}|\boldsymbol{x}_0) - p(t, \boldsymbol{y}|\boldsymbol{x}_1)\|_w^2 \\ &= \sum_{\boldsymbol{y}} (p(t, \boldsymbol{y}|\boldsymbol{x}_0) - p(t, \boldsymbol{y}|\boldsymbol{x}_1))^2 w(\boldsymbol{y}) \end{aligned} \tag{10.10}$$

with the specific choice $w(\boldsymbol{y}) = 1/\phi_0(\boldsymbol{y})$ for the weight function, which takes into account the (empirical) local density of the points, and puts more weight on low density points.

Since this distance depends on the random walk on the graph, we quite naturally denote it as the *diffusion distance* at time t. We also denote the mapping between the original space and the first k eigenvectors as the *diffusion map* at time t

$$\Psi_t(\boldsymbol{x}) = \left(\lambda_1^t \psi_1(\boldsymbol{x}), \lambda_2^t \psi_2(\boldsymbol{x}), \ldots, \lambda_k^t \psi_k(\boldsymbol{x})\right) \ . \tag{10.11}$$

The following theorem relates the diffusion distance and the diffusion map.
Theorem: *The diffusion distance (10.10) is equal to Euclidean distance in the diffusion map space with all* $(n-1)$ *eigenvectors.*

$$D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) = \sum_{j\geq 1} \lambda_j^{2t} \left(\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1)\right)^2 = \|\Psi_t(\boldsymbol{x}_0) - \Psi_t(\boldsymbol{x}_1)\|^2 \ . \tag{10.12}$$

Proof: Combining (10.9) and (10.10) gives

$$D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) = \sum_{\boldsymbol{y}} \Big(\sum_j \lambda_j^t (\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1)) \phi_j(\boldsymbol{y}) \Big)^2 / \phi_0(\boldsymbol{y}) \; . \tag{10.13}$$

Expanding the brackets and changing the order of summation gives

$$\begin{aligned} &D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) \\ &\quad = \sum_{j,k} \lambda_j^t \left(\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1)\right) \lambda_k^t \left(\psi_k(\boldsymbol{x}_0) - \psi_k(\boldsymbol{x}_1)\right) \sum_{\boldsymbol{y}} \frac{\phi_j(\boldsymbol{y}) \phi_k(\boldsymbol{y})}{\phi_0(\boldsymbol{y})} \; . \end{aligned}$$

From relation (10.4) it follows that $\phi_k / \phi_0 = \psi_k$. Moreover, according to (10.5) the vectors ϕ_j and ψ_k are bi-orthonormal. Therefore, the inner summation over $\boldsymbol{y}$ gives δ_{jk}, and overall the required formula (10.12). Note that in (10.12) summation starts from $j \geq 1$ since $\psi_0(\boldsymbol{x}) = 1$. □

This theorem provides a probabilistic interpretation to the non-linear embedding of points $\boldsymbol{x}_i$ from the original space (say $\mathbb{R}^p$) to the diffusion map space $\mathbb{R}^{n-1}$. Therefore, geometry in diffusion space is meaningful, and can be interpreted in terms of the Markov chain. The advantage of this distance measure over the standard distance between points in the original space is clear. While the original distance between any pair of points is independent of the location of all other points in the dataset, the diffusion distance between a pair of points depends on all possible paths connecting them, including those that pass through other points in the dataset. The diffusion distance thus measures the dynamical proximity between points on the graph, according to their connectivity.

Both the diffusion distance and the diffusion map depend on the time parameter t. For very short times, all points in the diffusion map space are far apart, whereas as time increases to infinity, all pairwise distances converge to zero, since $p(t, \boldsymbol{y}|\boldsymbol{x})$ converges to the stationary distribution. It is in the intermediate regime, where at different times different structures of the graph are revealed [11].

The identity (10.12) shows that the eigenvalues and eigenvectors $\{\lambda_j, \psi_j\}_{j \geq 1}$ capture the characteristic relaxation times and processes of the random walk on the graph. On a connected graph with n points, there are $n-1$ possible time scales. However, most of them capture fine detail structure and only the first few largest eigenvalues capture the coarse global structures of the graph. In cases where the matrix M has a *spectral gap* with only a few eigenvalues close to one and all remaining eigenvalues much smaller than one, the diffusion distance at a large enough time t can be well approximated by only the first few k eigenvectors $\psi_1(\boldsymbol{x}), \ldots, \psi_k(\boldsymbol{x})$, with a negligible error. Furthermore, as shown in [22], quantizing this diffusion space is equivalent to lumping the random walk, retaining only its slowest relaxation processes. The following lemma bounds the error of a k-term approximation of the diffusion distance.

Lemma: *For all times $t \geq 0$, the error in a k-term approximation of the diffusion distance is bounded by*

$$|D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) - \sum_{j=1}^{k} \lambda_j^{2t} (\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1))^2| \leq \lambda_{k+1}^{2t} \left(\frac{1}{\phi_0(\boldsymbol{x}_0)} + \frac{1}{\phi_0(\boldsymbol{x}_1)} \right) . \tag{10.14}$$

Proof: From the spectral decomposition (10.12)

$$\begin{aligned} |D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) - \sum_{j=1}^{k} \lambda_j^{2t} (\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1))^2| &= \sum_{j=k+1}^{n-1} \lambda_j^{2t} (\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1))^2 \\ &\leq \lambda_{k+1}^{2t} \sum_{j=0}^{n-1} (\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1))^2 . \end{aligned} \tag{10.15}$$

In addition, at time $t = 0$, we get that

$$D_0^2(\boldsymbol{x}_0, \boldsymbol{x}_1) = \sum_{j=0}^{n-1} (\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1))^2 .$$

However, from the definition of the diffusion distance (10.10), we have that at time $t = 0$

$$D_0^2(\boldsymbol{x}_0, \boldsymbol{x}_1) = \|p(0, \boldsymbol{y}|\boldsymbol{x}_0) - p(0, \boldsymbol{y}|\boldsymbol{x}_1)\|_w^2 = \left(\frac{1}{\phi_0(\boldsymbol{x}_0)} + \frac{1}{\phi_0(\boldsymbol{x}_1)} \right) .$$

Combining the last three equations proves the lemma. □.

Remark: This lemma shows that the error in computing an approximate diffusion distance with only k eigenvectors decays exponentially fast as a function of time. As the number of points $n \to \infty$, Eq. (10.14) is not informative since the steady state probabilities of individual points decay to zero at least as fast as $1/n$. However, for a very large number of points it makes more sense to consider the diffusion distance between regions of space instead of between individual points. Let Ω_1, Ω_2 be two such subsets of points. We then define

$$D_t^2(\Omega_1, \Omega_2) = \sum_{\boldsymbol{x}} \frac{(p(\boldsymbol{x}, t|\Omega_1) - p(\boldsymbol{x}, t|\Omega_2))^2}{\phi_0(\boldsymbol{x})} , \tag{10.16}$$

where $p(\boldsymbol{x}, t|\Omega_1)$ is the transition probability at time t, starting from the region Ω_1. As initial conditions inside Ω_i, we choose the steady state distribution, conditional on the random walk starting inside this region,

$$p(\boldsymbol{x}, 0|\Omega_i) = p_i(\boldsymbol{x}) = \begin{cases} \dfrac{\phi_0(\boldsymbol{x})}{\phi_0(\Omega_i)} , & \text{if } \boldsymbol{x} \in \Omega_i ; \\ 0 , & \text{if } \boldsymbol{x} \notin \Omega_i , \end{cases} \tag{10.17}$$

where

$$\phi_0(\Omega_i) = \sum_{\boldsymbol{y}\in\ \Omega_i} \phi_0(\boldsymbol{y}) \ . \tag{10.18}$$

Eq. (10.16) can then be written as

$$D_t^2(\Omega_1,\Omega_2) = \sum_j \lambda_j^{2t} \left(\psi_j(\Omega_1) - \psi_j(\Omega_2)\right)^2 \ , \tag{10.19}$$

where $\psi_j(\Omega_i) = \sum_{\boldsymbol{x}\in\Omega_i} \psi_j(\boldsymbol{x}) p_i(\boldsymbol{x})$. Similar to the proof of the lemma, it follows that

$$|D_t^2(\Omega_1,\Omega_2) - \sum_{j=0}^{k} \lambda_j^{2t}(\psi_j(\Omega_1) - \psi_j(\Omega_2))^2| \le \lambda_{k+1}^{2t} \left[\frac{1}{\phi_0(\Omega_1)} + \frac{1}{\phi_0(\Omega_2)}\right] \ . \tag{10.20}$$

Therefore, if we take regions Ω_i with non negligible steady state probabilities that are bounded from below by some constant, $\phi_0(\Omega_i) > \alpha$, for times $t \gg |\log(\lambda_{k+1})/\log(\alpha)|$, the approximation error of the k-term expansion is negligible. This observation provides a probabilistic interpretation as to what information is lost and retained in dimensional reduction with these eigenvectors.

In addition, the following theorem shows that this k-dimensional approximation is *optimal* under a certain mean squared error criterion.

Theorem: *Out of all k-dimensional approximations of the form*

$$\hat{p}_k(t,\boldsymbol{y}|\boldsymbol{x}) = \phi_0(\boldsymbol{y}) + \sum_{j=1}^{k} a_j(t,\boldsymbol{x}) \boldsymbol{w}_j(\boldsymbol{y})$$

for the probability distribution at time t, the one that minimizes the mean squared error

$$\mathbb{E}_{\boldsymbol{x}}\{\|p(t,\boldsymbol{y}|\boldsymbol{x}) - \hat{p}_k(t,\boldsymbol{y}|\boldsymbol{x})\|_w^2\} \ ,$$

where averaging over initial points $\boldsymbol{x}$ is with respect to the stationary density $\phi_0(\boldsymbol{x})$, is given by $\boldsymbol{w}_j(\boldsymbol{y}) = \phi_j(\boldsymbol{y})$ and $a_j(t,\boldsymbol{x}) = \lambda_j^t \psi_j(\boldsymbol{x})$. Therefore, the optimal k-dimensional approximation is given by the truncated sum

$$\hat{p}_k(\boldsymbol{y},t|\boldsymbol{x}) = \phi_0(\boldsymbol{y}) + \sum_{j=1}^{k} \lambda_j^t \psi_j(\boldsymbol{x}) \phi_j(\boldsymbol{y}) \ . \tag{10.21}$$

Proof: The proof is a consequence of weighted principal component analysis applied to the matrix M, taking into account the bi-orthogonality of the left and right eigenvectors.

We note that the first few eigenvectors are also optimal under other criteria, for example for data sampled from a manifold as in [6], or for multiclass spectral clustering [23].

10.2.1 Asymptotics of the Diffusion Map

Further insight into the properties of spectral clustering can be gained by considering the limit as the number of samples converges to infinity, and as the width of the kernel approaches zero. This has been the subject of intensive research over the past few years by various authors [6, 26, 29, 11, 24, 25, 27, 28]. Here we present the main results without detailed mathematical proofs and refer the reader to the above works.

The starting point for the analysis of this limit is the introduction of a statistical model in which the data points $\{\boldsymbol{x}_i\}$ are i.i.d. random samples from a smooth probability density $p(\boldsymbol{x})$ confined to a compact connected subset $\Omega \subset \mathbb{R}^p$ with smooth boundary $\partial\Omega$. Following the statistical physics notation, we write the density in Boltzmann form, $p(\boldsymbol{x}) = e^{-U(\boldsymbol{x})}$, where $U(\boldsymbol{x})$ is the (dimensionless) potential or energy of the configuration $\boldsymbol{x}$.

For each eigenvector $\boldsymbol{v}_j$ of the discrete matrix M with corresponding eigenvalue $\lambda_j \neq 0$, we extend its definition to any $\boldsymbol{x} \in \Omega$ as follows

$$\psi_j^{(n)}(\boldsymbol{x}) = \frac{1}{\lambda_j} \sum_i \frac{k(\boldsymbol{x}, \boldsymbol{x}_i)}{D(\boldsymbol{x})} \boldsymbol{v}_j(\boldsymbol{x}_i) \tag{10.22}$$

with $D(\boldsymbol{x}) = \sum_i k(\boldsymbol{x}, \boldsymbol{x}_i)$. Note that this definition retains the values at the sampled points, e.g., $\psi_j^{(n)}(\boldsymbol{x}_i) = v_j(\boldsymbol{x}_i)$ for all $i = 1, \ldots, n$.

As shown in [24], in the limit $n \to \infty$ the random walk on the discrete graph converges to a random walk on the continuous space Ω. Then, it is possible to define an integral operator T as follows,

$$T[\psi](\boldsymbol{x}) = \int_\Omega M(\boldsymbol{y}|\boldsymbol{x}) \psi(\boldsymbol{y}) p(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \ ,$$

where $M(\boldsymbol{x}|\boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/\sigma^2)/D_\sigma(\boldsymbol{y})$ is the transition probability from $\boldsymbol{y}$ to $\boldsymbol{x}$ in time ε, and $D_\sigma(\boldsymbol{y}) = \int_\Omega \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/\sigma^2) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$. In the limit $n \to \infty$, the eigenvalues λ_j and the extensions $\psi_j^{(n)}$ of the discrete eigenvectors v_j converge to the eigenvalues and eigenfunctions of the integral operator T.

Further, in the limit $\sigma \to 0$, the random walk on the space Ω, upon scaling of time, converges to a diffusion process in a potential $2U(\boldsymbol{x})$,

$$\dot{\boldsymbol{x}}(t) = -\nabla(2U) + \sqrt{2}\dot{\boldsymbol{w}}(t) \ , \tag{10.23}$$

where $U(\boldsymbol{x}) = -log(p(\boldsymbol{x}))$ and $\boldsymbol{w}(t)$ is standard Brownian motion in p dimensions. In this limit, the eigenfunctions of the integral operator T converge to those of the infinitesimal generator of this diffusion process, given by the following Fokker-Planck (FP) operator,

$$\mathcal{H}\psi = \Delta\psi - 2\nabla\psi \cdot \nabla U \ . \tag{10.24}$$

Table 10.1. Random Walks and Diffusion Processes

Case	Operator	Stochastic Process
$\sigma > 0$	finite $n \times n$	R.W. discrete in space
$n < \infty$	matrix M	discrete in time
$\sigma > 0$	integral	R.W. in continuous space
$n \to \infty$	operator T	discrete in time
$\sigma \to 0$	infinitesimal	diffusion process
$n \to \infty$	generator $\mathcal{H}$	continuous in time & space

The Langevin equation (10.23) is the standard model to describe stochastic dynamical systems in physics, chemistry and biology [30, 31]. As such, its characteristics as well as those of the corresponding FP equation have been extensively studied, see [30, 31, 32] and references therein. The term $\nabla\psi \cdot \nabla U$ in (10.24) is interpreted as a *drift* term towards low energy (high-density) regions, and plays a crucial part in the definition of clusters.

Note that when data is *uniformly sampled* from Ω, $\nabla U = 0$ so the drift term vanishes and we recover the Laplace-Beltrami operator on Ω.

Finally, when the density p has compact support on a domain Ω, the operator $\mathcal{H}$ is defined only inside Ω. Its eigenvalues and eigenfunctions thus depend on the boundary conditions at $\partial\Omega$. As shown in [11], in the limit $\sigma \to 0$ the random walk satisfies reflecting boundary conditions on $\partial\Omega$, which translate into

$$\left.\frac{\partial\psi(\boldsymbol{x})}{\partial \boldsymbol{n}}\right|_{\partial\Omega} = 0 \,, \tag{10.25}$$

where $\boldsymbol{n}$ is a unit normal vector at the point $\boldsymbol{x} \in \partial\Omega$.

To conclude, the right eigenvectors of the finite matrix M can be viewed as discrete approximations to those of the operator T, which in turn can be viewed as approximations to those of $\mathcal{H}$. Therefore, if there are enough data points for accurate statistical sampling, the structure and characteristics of the eigenvalues and eigenfunctions of $\mathcal{H}$ are similar to the corresponding eigenvalues and discrete eigenvectors of M. In the next sections we show how this relation can be used to explain the characteristics of spectral clustering and dimensional reduction algorithms. The three different stochastic processes are summarized in table 1.

10.3 Spectral Embedding of Low Dimensional Manifolds

Let $\{\boldsymbol{x}_i\}$ denote points sampled (uniformly for simplicity) from a low dimensional manifold embedded in a high dimensional space. Eq. (10.14) shows that by retaining only the first k coordinates of the diffusion map, the reconstruction error of the long time random walk transition probabilities is

negligible. However, this is not necessarily the correct criterion for an embedding algorithm. Broadly speaking, assuming that data is indeed sampled from a manifold, a low dimensional embedding should preserve (e.g. uncover) the information about the global (coarse) structure of this manifold, while throwing out information about its fine details. A crucial question is then under what conditions does spectral embedding indeed satisfy these requirements, and perhaps more generally, what are its characteristics.

The manifold learning problem of a low dimensional embedding can be formulated as follows: Let $\mathcal{Y} = \{\boldsymbol{y}_i\}_{i=1}^n \subset \mathbb{R}^q$ denote a set of points randomly sampled from some smooth probability density defined in a compact domain of $\mathbb{R}^q$ (the coordinate space). However, we are given the set of points $\mathcal{X} = f(\mathcal{Y})$ where $f : \mathbb{R}^q \to \mathbb{R}^p$ is a smooth mapping with $p > q$. Therefore, assuming that the points $\mathcal{Y}$ are not themselves on a lower dimensional manifold than $\mathbb{R}^q$, then the points $\mathcal{X}$ lie on a manifold of dimension q in $\mathbb{R}^p$. Given $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, the problem is to estimate the dimensionality q and the coordinate points $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$. Obviously, this problem is ill-posed and various degrees of freedom, such as translation, rotation, reflection and scaling cannot be determined.

While a general theory of manifold learning is not yet fully developed, in this section we would like to provide a glimpse into the properties of spectral embeddings, based on the probabilistic interpretation of section 10.2. We prove that in certain cases spectral embedding works, in the sense that it finds a reasonable embedding of the data, while in other cases modifications to the basic scheme are needed.

We start from the simplest example of a one dimensional curve embedded in a higher dimensional space. In this case, a successful low dimensional embedding should uncover the one-dimensionality of the data and give a representation of the arclength of the curve. We prove that spectral embedding succeeds in this task:

Theorem: *Consider data sampled* uniformly *from a non-intersecting smooth 1-D curve embedded in a high dimensional space. Then, in the limit of a large number of samples and small kernel width the first diffusion map coordinate gives a one-to-one parametrization of the curve. Further, in the case of a closed curve, the first two diffusion map coordinates map the curve into the circle.*

Proof: Let $\Gamma : [0, 1] \to \mathbb{R}^p$ denote a constant speed parametrization s of the curve ($\|d\Gamma(s)/ds\| = const$). As $n \to \infty, \varepsilon \to 0$, the diffusion map coordinates (eigenvectors of M) converge to the eigenfunctions of the corresponding FP operator. In the case of a non-intersecting 1-D curve, the Fokker-Planck operator is

$$\mathcal{H}\psi = \frac{d^2\psi}{ds^2} \,, \tag{10.26}$$

where s is an arc-length along Γ, with Neumann boundary conditions at the edges $s = 0, 1$. The first two non-trivial eigenfunctions are $\psi_1 = \cos(\pi s)$ and $\psi_2 = \cos(2\pi s)$. The first eigenfunction thus gives a one-to-one *parametrization* of the curve, and can thus be used to embed it into $\mathbb{R}^1$. The second eigenfunction $\psi_2 = 2\psi_1^2 - 1$ is a quadratic function of the first. This relation (together with estimates on the local density of the points) can be used to verify that for a given dataset, at a coarse scale its data points indeed lie on a 1-D manifold.

Consider now a *closed* curve in $\mathbb{R}^p$. In this case there are no boundary conditions for the operator and we seek periodic eigenfunctions. The first two non-constant eigenfunctions are $\sin(\pi s + \theta)$ and $\cos(\pi s + \theta)$ where θ is an arbitrary rotation angle. These two eigenfunctions map data points on the curve to the circle in $\mathbb{R}^2$, see [11]. □

Example 1: Consider a set of 400 points in three-dimensions, sampled uniformly from a spiral curve. In figure 10.1 the points and the first two eigenvectors are plotted. As expected, the first eigenvector provides a parametrization of the curve, whereas the second one is a quadratic function of the first.

Example 2: The analysis above can also be applied to images. Consider a dataset of images of a single object taken from different horizontal rotation angles. These images, although residing in a high dimensional space, are all on a 1-d manifold defined by the rotation angle. The diffusion map can uncover this underlying one dimensional manifold on which the images reside and organize the images according to it. An example is shown in figure 10.2, where the first two diffusion map coordinates computed on a dataset of 37 images of a truck taken at uniform angles of $0, 5, \ldots, 175, 180$ degrees are plotted one against the other. All computations were done using a Gaussian kernel with standard Euclidean distance between all images. The data is courtesy of Ronen Basri [33].

We remark that if data is sampled from a 1-D curve or more generally from a low dimensional manifold, but not in a uniform manner, the standard normalized graph Laplacian converges to the FP operator (10.24) which contains a drift term. Therefore its eigenfunctions depend both on the geometry of the manifold and on the probability density on it. However, replacing the isotropic kernel $\exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/4\varepsilon)$ by the anisotropic one $\exp(-\|\boldsymbol{x} - \boldsymbol{y}\|/4\varepsilon)/D(\boldsymbol{x})D(\boldsymbol{y})$ asymptotically removes the effects of density and retains only those of geometry. With this kernel, the normalized graph Laplacian converges to the Laplace-Beltrami operator on the manifold [11].

We now consider the characteristics of spectral embedding on the "swiss roll" dataset, which has been used as a synthetic benchmark in many papers, see [7, 34] and refs. therein. The swiss roll is a 2-D manifold embedded in $\mathbb{R}^3$. A set of n points $\boldsymbol{x}_i \in \mathbb{R}^3$ are generated according to $\boldsymbol{x} = (t\cos(t), h, t\sin(t))$, where $t \sim U[3\pi/2, 9\pi/2]$, and $h \sim U[0, H]$. By unfolding the roll, we obtain a rectangle of length L and width H, where in our example,

$$L = \int_{3\pi/2}^{9\pi/2} \sqrt{\left(\frac{\mathrm{d}}{\mathrm{d}t} t\sin t\right)^2 + \left(\frac{\mathrm{d}}{\mathrm{d}t} tcos(t)\right)^2}\, \mathrm{d}t \approx 90\ .$$

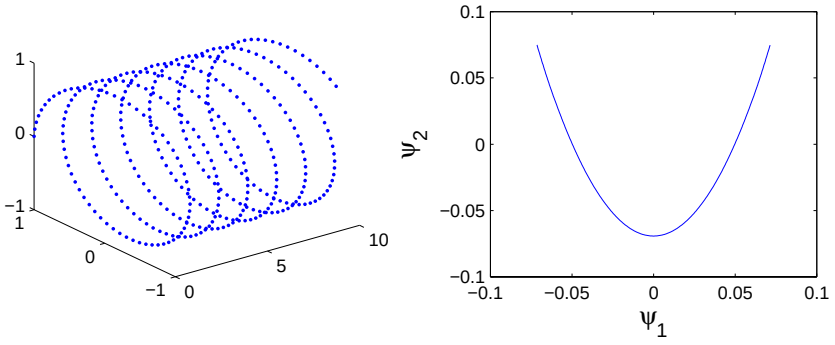

Fig. 10.1. 400 points uniformly sampled from a spiral in 3-D (left). First two non-trivial eigenfunctions. The first eigenfunction ψ_1 provides a parametrization of the curve. The second one is a quadratic function of the first

For points uniformly distributed on this manifold, in the limit $n \to \infty, \varepsilon \to 0$, the FP operator is

$$\mathcal{H}\psi = \frac{d^2\psi}{dt^2} + \frac{d^2\psi}{dh^2}$$

with Neumann boundary conditions at the boundaries of the rectangle. Its eigenvalues and eigenfunctions are

$$\begin{aligned} \mu_{j,k} &= \pi^2 \left(\frac{j^2}{L^2} + \frac{k^2}{H^2} \right), \quad j,k \geq 0 \; ; \\ \psi(t,h) &= \cos\left(\frac{j\pi t}{L} \right) \cos\left(\frac{k\pi h}{H} \right) \; . \end{aligned} \tag{10.27}$$

First we consider a reasonably wide swiss roll, with $H = 50$. In this case, the length and width of the roll are similar and so upon ordering the eigenvalues $\mu_{j,k}$ in increasing order, the first two eigenfunctions after the constant one are $\cos(\pi t/L)$ and $\cos(\pi h/H)$. In this case spectral embedding via the first two diffusion map coordinates gives a reasonably nice parametrization of the manifold, uncovering its 2-d nature, see fig. 10.3.

However, consider now the same swiss roll but with a slightly smaller width $H = 30$. Now the roll is roughly three times as long as it is wide. In this case, the first eigenfunction $\cos(\pi t/L)$ gives a one-to-one parametrization of the parameter t. However, the next two eigenfunctions, $\cos(2\pi t/L)$ and $\cos(3\pi t/L)$, are functions of ψ_1, and thus provide no further useful information for the low dimensional representation of the manifold. It is only the 4th eigenfunction that reveals its two dimensional nature, see fig. 10.4. We remark that in both figures we do not obtain perfect rectangles in the embedded space. This is due to the non-uniform density of points on the manifold, with points more densely sampled in the inward spiral than in the outward one.

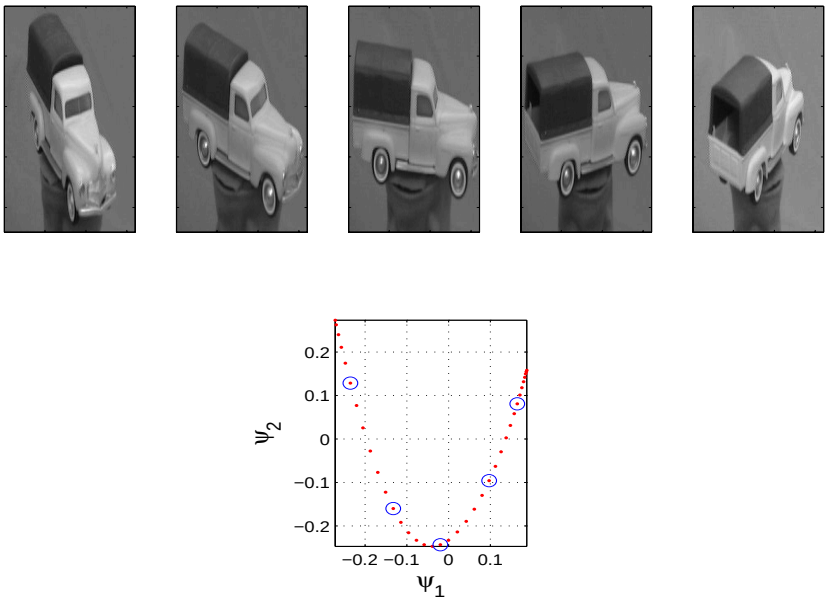

Fig. 10.2. Figures of a truck taken at five different horizontal angles (top). The mapping of the 37 images into the first two eigenvectors, based on a Gaussian kernel with standard Euclidean distance between the images as the underlying metric (bottom). The blue circles correspond to the five specific images shown above

This example shows a fundamental difference between (linear) low dimensional embedding by principal component analysis, vs. nonlinear spectral methods. In PCA once the variance in a specific direction has been captured, all further projections are orthogonal to it. In non-linear spectral methods, the situation is fundamentally different. For example, even for points on a one dimensional (linear) line segment, there are N different eigenvectors that capture the various relaxation processes on it, all with non-zero eigenvalues. Therefore, several eigenvectors may encode for the same geometrical or spatial "direction" of a manifold. To obtain a sensible low dimensional representation, an analysis of the relations between the different eigenvectors is required to remove this redundancy.

10.4 Spectral Clustering of a Mixture of Gaussians

A second common application of spectral embedding methods is for the purpose of clustering. Given a set of n points $\{\boldsymbol{x}_i\}_{i=1}^n$ and a corresponding similarity matrix W_{ij}, many works suggest to use the first few coordinates of the normalized graph Laplacian as an embedding into a new space, where standard clustering algorithms such as k-means can be employed. Most methods

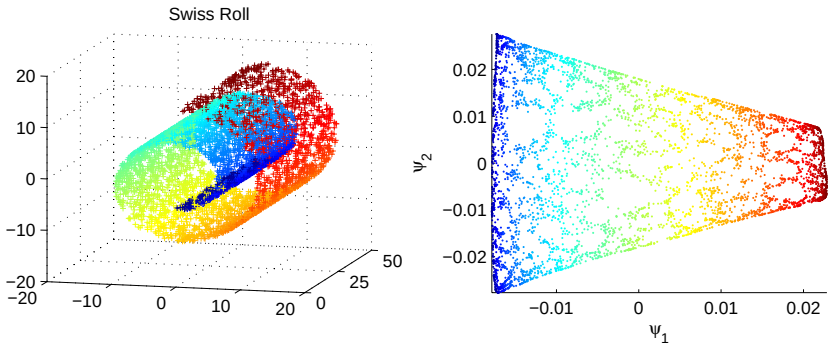

Fig. 10.3. 5000 points sampled from a wide swiss roll and embedding into the first two diffusion map coordinates

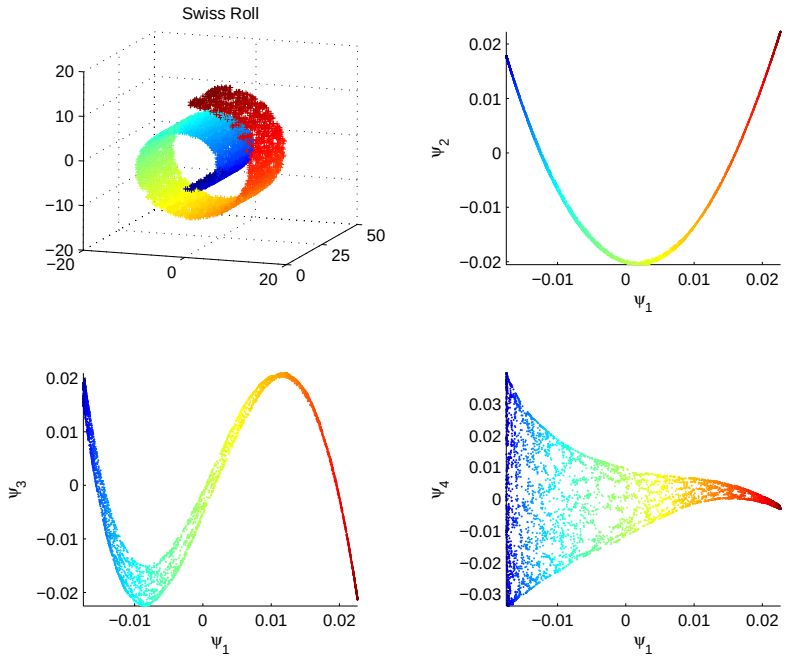

Fig. 10.4. 5000 points sampled from a narrow swiss roll and embedding into various diffusion map coordinates

suggest to use the first $k-1$ non-trivial eigenvectors after the constant one to find k clusters in a dataset. The various methods differ by the exact normalization of the matrix for which the eigenvectors are computed and the specific clustering algorithm applied after the embedding into the new space. Note

that if the original space had dimensionality $p < k$, then the embedding actually *increases* the dimension of the data for clustering purposes. An interesting question is then under what conditions are these spectral embedding followed by standard clustering methods expected to yield successful clustering results.

Two ingredients are needed to analyze this question. The first is a generative model for clustered data, and the second is an explicit definition of what is considered a good clustering result.

A standard generative model for data in general and for clustered data in particular is the *mixture of Gaussians* model. In this setting, data points $\{\boldsymbol{x}_i\}$ are i.i.d. samples from a density composed of a mixture of K Gaussians,

$$p(\boldsymbol{x}) = \sum_{i=1}^{K} w_i N(\boldsymbol{\mu}_i, \Sigma_i) \tag{10.28}$$

with means $\boldsymbol{\mu}_i$, covariance matrices Σ_i and respective weights w_i. We say that data from such a model is clusterable into K clusters if all the different Gaussian clouds are well separated from each other. This can be translated into the condition that

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 > 2\min[\lambda_{max}(\Sigma_i), \lambda_{max}(\Sigma_j)] \qquad \forall i, j\ , \tag{10.29}$$

where $\lambda_{max}(\Sigma)$ is the largest eigenvalue of a covariance matrix Σ.

Let $\{\boldsymbol{x}_i\}$ denote a dataset from a mixture that satisfies these conditions, and let $S_1 \cup S_2 \cup \ldots \cup S_K$ denote the partition of space into K disjoint regions, where each region S_j is defined to contain all points $\boldsymbol{x} \in \mathbb{R}^p$ whose probability to have been sampled from the j-th Gaussian is the largest. We consider the output of a clustering algorithm to be successful if its K regions have a high overlap to these optimal Bayes regions S_j.

We now analyze the performance of spectral clustering in this setting. We assume that we have a very large number of points and do not consider the important issue of finite sample size effects. Furthermore, we do not consider a specific spectral clustering algorithm, but rather give general statements regarding their possible success given the structure of the embedding coordinates.

In our analysis, we employ the intimate connection between the diffusion distance and the characteristic time scales and relaxation processes of the random walk on the graph of points, combined with matrix perturbation theory. A similar analysis can be made using the properties of the eigenvalues and eigenfunctions of the limiting FP operator.

Consider then n data points $\{\boldsymbol{x}_i\}_{i=1}^n$ sampled from a mixture of K reasonably separated Gaussians, and let $S_1 \cup S_2 \cup \ldots S_K$ denote a partition of space into K disjoint cluster regions as defined above. Then, by definition, each cluster region S_j contains the majority of points of each respective Gaussian. Consider the similarity matrix W computed on this discrete dataset, where we sort the points according to which cluster region they belong to. Since the

Gaussians are partially overlapping, the similarity matrix W does not have a perfect block structure (with the blocks being the sets S_j), but rather has small non zero weights between points of different cluster regions. To analyze the possible behavior of the eigenvalues and eigenvectors of such matrices, we introduce the following quantities. For each point $\boldsymbol{x}_i \in S_j$ we define

$$a_i = \sum_{\boldsymbol{x}_k \notin S_j} W_{ik} \tag{10.30}$$

and

$$b_i = \sum_{\boldsymbol{x}_k \in S_j} W_{ik} \ . \tag{10.31}$$

The quantity a_i measures the amount of connectivity of the point $\boldsymbol{x}_i$ to points outside its cluster, whereas b_i measures the amount of connectivity to points in the same cluster. Further, we introduce a family of similarity matrices depending on a parameter ε, as follows:

$$W(\varepsilon) = (1-\varepsilon) diag\left(\frac{a_i}{b_i}\right) W_0 + \varepsilon W_1 \ , \tag{10.32}$$

where

$$W_0(i,j) = \begin{cases} W_{ij} \ , & \text{if } \boldsymbol{x}_i \ , \boldsymbol{x}_j \in S_k \ , i \neq j \ ; \\ 0 \ , & \text{otherwise} \ , \end{cases} \tag{10.33}$$

and

$$W_1(i,j) = \begin{cases} W_{ij} \ , & \text{if } \boldsymbol{x}_i \in S_\alpha \ , \boldsymbol{x}_j \in S_\beta \ , \alpha \neq \beta \ ; \\ 0 \ , & \text{otherwise} \ . \end{cases} \tag{10.34}$$

The matrix W_0 is therefore a block matrix with K blocks, which contains all intra-cluster connections, while the matrix W_1 contains all the inter-cluster connections. Note that in the representation (10.32), for each point $\boldsymbol{x}_i$, $D(\boldsymbol{x}_i) = \sum W_{ij}(\varepsilon)$ is independent of ε. Therefore, for the symmetric matrix $M_s(\varepsilon)$ similar to the Markov matrix, we can write

$$M_s(\varepsilon) = D^{1/2} W(\varepsilon) D^{1/2} = M_s(0) + \varepsilon M_1 \ . \tag{10.35}$$

When $\varepsilon = 0$, $W(\varepsilon) = W_0$ is a block matrix and so the matrix $M_s(0)$ corresponds to a *reducible* Markov chain with K components. When $\varepsilon = 1$ we obtain the original Markov matrix on the dataset, whose eigenvectors will be used to cluster the data. The parameter ε can thus be viewed as controlling the strength of the inter-cluster connections. Our aim is to relate the eigenvalues and eigenvectors of $M_s(0)$ to those of $M_s(1)$, while viewing the matrix εM_1 as a small perturbation.

Since $M_s(0)$ corresponds to a Markov chain with K disconnected components, the eigenvalue $\lambda = 1$ has multiplicity K. Further, we denote by $\lambda_1^R, \ldots, \lambda_K^R$ the next largest eigenvalue in each of the K blocks. These eigenvalues correspond to the characteristic relaxation times in each of the K clusters, (denoted as spurious eigenvalues in [14]). As ε is increased from zero, the

eigenvalue $\lambda = 1$ with multiplicity K splits into K different branches. Since $M_s(\varepsilon)$ is a Markov matrix for all $0 \leq \varepsilon \leq 1$ and becomes connected for $\varepsilon > 0$, exactly one of the K eigenvalues stays fixed at $\lambda = 1$, whereas the remaining $K-1$ decrease below one. These slightly smaller than one eigenvalues capture the mean exit times from the now weakly connected clusters.

According to Kato [35], [Theorem 6.1, page 120], the eigenvalues and eigenvectors of $M(\varepsilon)$ are *analytic* functions of ε on the real line. The point $\varepsilon = 0$, where $\lambda = 1$ has multiplicity $K > 1$ is called an *exceptional point*. Further, (see Kato [35], page 124) if we sort the eigenvalues in decreasing order, then the graph of each eigenvalue as a function of ε is a continuous function, which may cross other eigenvalues at various exceptional points ε_j, At each one of these values of ε, the graph of the eigenvalue as a function of ε jumps from one smooth curve to another. The corresponding eigenvectors, however, change abruptly at these crossing points as they move from one eigenvector to a different one.

We now relate these results to spectral clustering. A set of points is considered clusterable by these spectral methods if the corresponding perturbation matrix M_1 is small, that is, if there are no exceptional points or eigenvalue crossings for all values $\varepsilon \in (0, 1)$. This means that the fastest exit time from either one of the clusters is significantly slower than the slowest relaxation time in each one of the clusters. In this case, the first $K - 1$ eigenvectors of the Markov matrix M are approximately piecewise constant inside each of the K clusters. The next eigenvectors capture relaxation processes inside individual clusters and so each of them is approximately zero in all clusters but one. Due to their weighted bi-orthogonality of all eigenvectors (see section 10.2), clustering the points according to the sign structure of the first $K - 1$ eigenvectors approximately recovers the K clusters. This is the setting in which we expect spectral clustering algorithms to succeed.

However, now consider the case where relaxation times of some clusters are larger than the mean exit times from other clusters. Then there exists at least one exceptional point $\varepsilon < 1$, where a crossing of eigenvalues occurs. In this case, crucial information required for successful clustering is lost in the first $K-1$ eigenvectors, since at least one of them now captures the relaxation process inside a large cluster. In this case, regardless of the specific clustering algorithm employed on these spectral embedding coordinates, it is not possible to distinguish one of the small clusters from others.

Example: We illustrate the results of this analysis on a simple example. Consider $n = 1000$ points generated from a mixture of three Gaussians in two dimensions. The centers of the Gaussians are

$$\mu_1 = (-6, 0),\ \mu_2 = (0, 0),\ \mu_3 = (x_R, 0)\ ,$$

where x_R is a parameter. The two rightmost Gaussians are spherical with standard deviation $\sigma_2 = \sigma_3 = 0.5$. The leftmost cluster has a diagonal covariance matrix

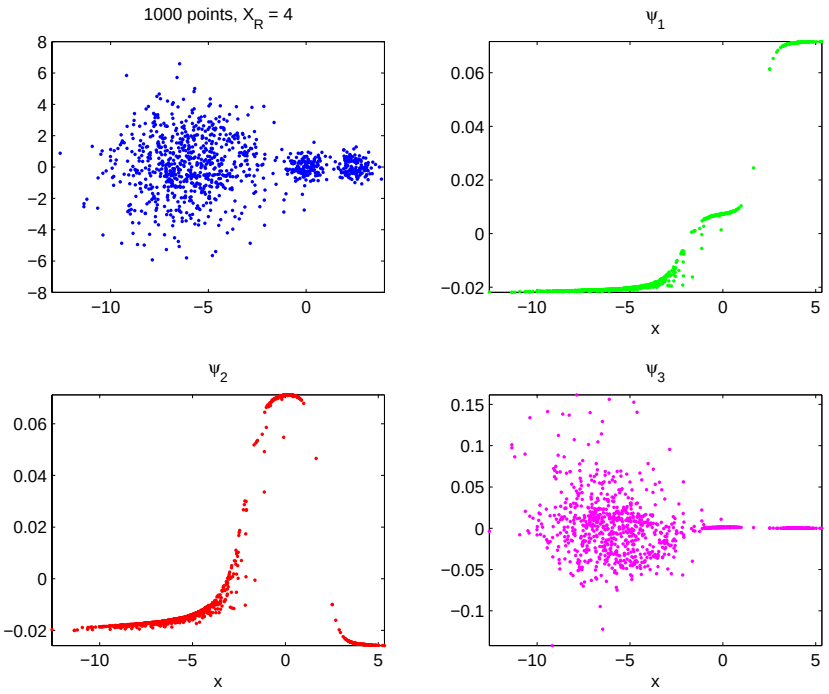

Fig. 10.5. Top left - 1000 points from three Gaussians. The three other panels show the first three non-trivial eigenvectors as a function of the x-coordinate

$$\Sigma_1 = \begin{pmatrix} 2.0 & 0 \\ 0 & 2.4 \end{pmatrix} .$$

The weights of the three clusters are $(w_1, w_2, w_3) = (0.7, 0.15, 0.15)$. In figure 10.5 we present the dataset of 1000 points sampled from this mixture with $x_R = 4$, and the resulting first three non-trivial eigenvectors, ψ_1, ψ_2, ψ_3 as a function of the x-axis. All computations were done with a Gaussian kernel with width $\sigma = 1.0$. As seen in the figure, the three clusters are well separated and thus the first two non-trivial eigenvectors are piecewise constant in each cluster, while the third eigenvector captures the relaxation along the y-axis in the leftmost Gaussian and is thus not a function of the x-coordinate. We expect spectral clustering that uses only ψ_1 and ψ_2 to succeed in this case.

Now consider a very similar dataset, only that the center x_R of the rightmost cluster is slowly decreased from $x_R = 4$ towards $x = 0$. The dependence of the top six eigenvalues on x_R is shown in figure 10.6. As seen from the top panel, the first eigenvalue crossing occurs at the exceptional point $x_R = 2.65$, and then additional crossings occur at $x_R = 2.4, 2.3$ and at 2.15.

Therefore, as long as $x_R > 2.65$ the mean exit time from the rightmost cluster is slower than the relaxation time in the large cluster, and spectral

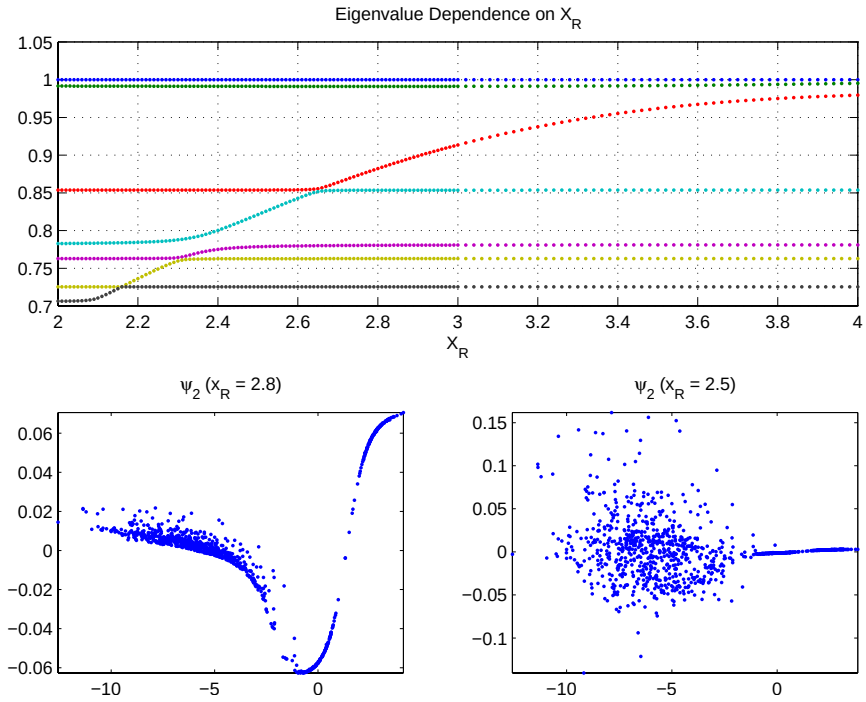

Fig. 10.6. Dependence of six largest eigenvalues on location of right cluster center (Top). The second largest non-trivial eigenvector as a function of the x-coordinate when $x_R = 2.8$ (Bottom left) and when $x_R = 2.5$ (Bottom right)

clustering using ψ_1, ψ_2 should be successful. However, for $x_R < 2.65$ the information distinguishing the two small clusters is not present any more in ψ_1, ψ_2 and thus spectral clustering will not be able to distinguish between these two clusters. An example of this sharp transition in the shape of the second eigenvector ψ_2 is shown in fig. 10.6 at the bottom left and right panels. For $x_R = 2.8 > 2.65$ the second eigenvector is approximately piecewise constant with two different constants in the two small clusters, whereas for $x_R = 2.5 < 2.65$ the second eigenvector captures the relaxation process in the large cluster and is approximately zero on both of the small ones. In this case ψ_3 captures the difference between these two smaller clusters.

As x_R is decreased further, additional eigenvalue crossings occur. In fig. 10.7 we show the first five non-trivial eigenvectors as a function of the x-coordinate for $x_R = 2.25$. Here, due to multiple eigenvalue crossings only ψ_5 is able to distinguish between the two rightmost Gaussians.

Our analysis shows that while spectral clustering may not work on multiscale data, the comparison of relaxation times inside one set of points vs. the mean first passage time between two sets of points plays a natural role in the

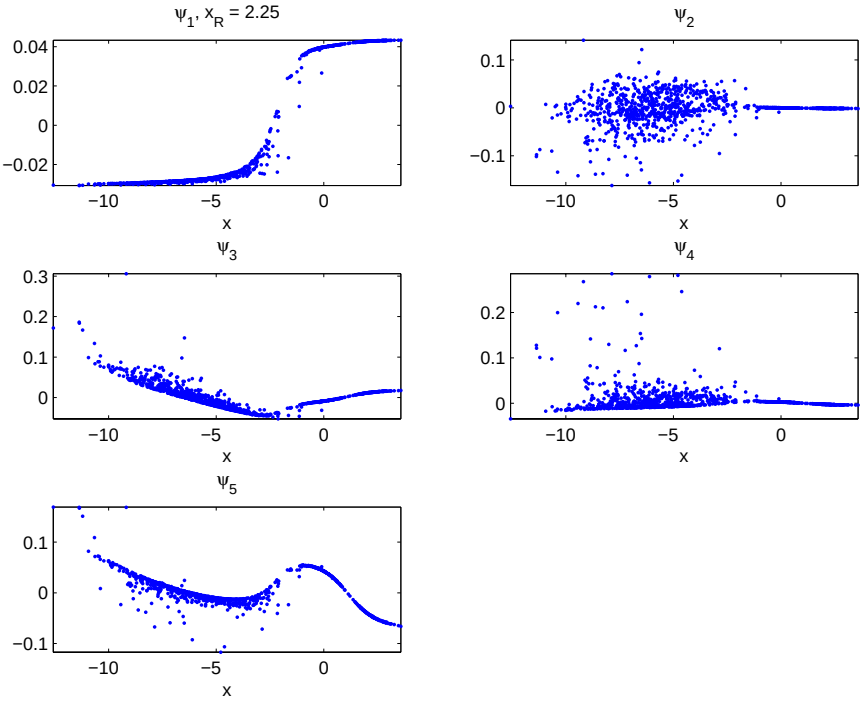

Fig. 10.7. The first five non-trivial eigenvectors as a function of the x-coordinate when the rightmost cluster is centered at $x_R = 2.25$

definition of clusters. This leads to a multi-scale approach to clustering, based on a relaxation time coherence measure for the determination of the coherence of a group of points as all belonging to a single cluster, see [36]. Such an approach is able to successfully cluster this example even when $x_R = 2.25$, and has also been applied to image segmentation problems.

Finally, we would like to mention a simple analogy between spectral clustering where the goal is the uncovering of clusters, and the uncovering of signals in (linear) principal component analysis. Consider a setting where we are given n observations of the type "signals + noise". A standard method to detect the signals is to compute the covariance matrix C of the observations and project the observations onto the first few leading eigenvectors of C. In this setting, if the signals lie in a low dimensional hyperspace of dimension k, and the noise has variance smaller than the smallest variance of the signals in this subspace, then PCA is successful at recovery of the signals. If, however, noise has variance larger than the smallest variance in this subspace, then at least one of the first k eigenvectors points in a direction orthogonal from this subspace, dictated by the direction with largest noise variance, and it is not possible to uncover all signals by PCA. Furthermore there is a sharp transi-

tion in the direction of this eigenvector, as noise strength is increased between being smaller than signal strength to larger than it [37]. As described above, in our case a similar sharp *phase transition phenomenon* occurs, only that the signal and the noise are replaced by other quantities: The "signals" are the mean exit times from the individual clusters, while the "noises" are the mean relaxation times inside them.

10.5 Summary and Discussion

In this paper we presented a probabilistic interpretation of spectral clustering and dimensionality reduction algorithms. We showed that the mapping of points from the feature space to the diffusion map space of eigenvectors of the normalized graph Laplacian has a well defined probabilistic meaning in terms of the diffusion distance. This distance, in turn, depends on both the geometry and density of the dataset. The key concepts in the analysis of these methods, that incorporates the density and geometry of a dataset, are the characteristic relaxation times and processes of the random walk on the graph. This provides novel insight into spectral clustering algorithms, and the starting point for the development of multiscale algorithms [36]. A similar analysis can also be applied to semi-supervised learning based on spectral methods [38]. Finally, these eigenvectors may be used to design better search and data collection protocols [39].

Acknowledgement. This work was partially supported by DARPA through AFOSR, and by the US department of Energy, CMPD (IGK). The research of BN is supported by the Israel Science Foundation (grant 432/06) and by the Hana and Julius Rosen fund.

References

1. Schölkopf, B. and Smola, A. J., and Müller, K.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, **10** (5), 1299–1319 (1998)
2. Weiss, Y.: Segmentation using eigenvectors: a unifying view. ICCV (1999)
3. Shi, J. and Malik, J.: Normalized cuts and image segmentation. PAMI, **22** (8), 888-905, (2000)
4. Ding, C., He, X., Zha, H., Gu, M., and Simon, H.: A min-max cut algorithm for graph partitioning and data clustering. In: Proc. IEEE International Conf. Data Mining, 107–114, (2001)
5. Cristianini, N., Shawe-Taylor, J., and Kandola, J.: Spectral kernel methods for clustering. NIPS, **14** (2002)
6. Belkin, M. and Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. NIPS, **14** (2002)

7. Belkin, M. and Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15**, 1373–1396 (2003)
8. Ng, A.Y., Jordan, M., and Weiss, Y.: On spectral clustering, analysis and an algorithm. NIPS, **14** (2002)
9. Zhu, X., Ghahramani, Z., and Lafferty J.: Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 20th international conference on machine learning (2003)
10. Saerens, M., Fouss, F., Yen L., and Dupont, P.: The principal component analysis of a graph and its relationships to spectral clustering. In: Proceedings of the 15th European Conference on Machine Learning, ECML, 371–383 (2004)
11. Coifman, R.R., Lafon, S.: Diffusion Maps. Appl. Comp. Harm. Anal., **21**, 5–30 (2006)
12. Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., and Zucker S.: Geometric diffusion as a tool for harmonic analysis and structure definition of data, parts I and II. Proc. Nat. Acad. Sci., **102** (21), 7426–7437 (2005)
13. Berard, P., Besson, G., Gallot, S.: Embedding Riemannian manifolds by their heat kernel. Geometric and Functional Analysis, **4** (1994)
14. Meila, M., Shi, J.: A random walks view of spectral segmentation. AI and Statistics (2001)
15. Yen, L., Vanvyve, D., Wouters, F., Fouss, F., Verleysen M., and Saerens, M.: Clustering using a random-walk based distance measure. In: Proceedings of the 13th Symposium on Artificial Neural Networks, ESANN, 317–324 (2005)
16. Tishby, N. and Slonim, N.: Data Clustering by Markovian Relaxation and the information bottleneck method. NIPS (2000)
17. Chennubhotla, C. and Jepson, A.J.: Half-lives of eigenflows for spectral clustering. NIPS (2002)
18. Harel, D. and Koren, Y.: Clustering spatial data using random walks. In: Proceedings of the 7th ACM Int. Conference on Knowledge Discovery and Data Mining, 281–286. ACM Press (2001)
19. Pons, P. and Latapy, M.: Computing Communities in Large Networks Using Random Walks. In: 20th International Symposium on Computer and Information Sciences (ISCIS'05). LNCS 3733 (2005)
20. Nadler, B., Lafon, S., Coifman, R.R., and Kevrekidis, I.G.: Diffusion maps spectral clustering and eigenfunctions of Fokker-Planck operators. NIPS (2005)
21. Parzen, E.: On estimation of a probability density function and mode. Ann. Math. Stat. **33**, 1065–1076 (1962)
22. Lafon, S. and Lee, A.B.: Diffusion maps: A unified framework for dimension reduction, data partitioning and graph subsampling. IEEE Trans. Patt. Anal. Mach. Int., **28** (9), 1393–1403 (2006)
23. Yu, S. and Shi, J.: Multiclass spectral clustering. ICCV (2003)
24. Nadler, B., Lafon, S., Coifman, R.R., and Kevrekidis, I.G.: Diffusion maps, spectral clustering, and the reaction coordinates of dynamical systems. Appl. Comp. Harm. Anal., **21**, 113–127 (2006)
25. von Luxburg, U., Bousquet, O., and Belkin, M.: On the convergence of spectral clustering on random samples: the normalized case. NIPS (2004)
26. Belkin, M. and Niyogi, P.: Towards a theoeretical foundation for Laplacian-based manifold methods. COLT (2005)
27. Hein, M., Audibert, J., and von Luxburg, U.: From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. COLT (2005)

28. Singer, A.: From graph to manifold Laplacian: the convergence rate. Applied and Computational Harmonic Analysis, **21** (1), 135–144 (2006)
29. Belkin, M. and Niyogi, P.: Convergence of Laplacian eigenmaps. NIPS (2006)
30. Gardiner, C.W.: Handbook of Stochastic Methods, 3rd edition. Springer, NY (2004)
31. Risken, H.: The Fokker Planck equation, 2nd edition. Springer NY (1999)
32. Matkowsky, B.J. and Schuss, Z.: Eigenvalues of the Fokker-Planck operator and the approach to equilibrium for diffusions in potential fields. SIAM J. App. Math. **40** (2), 242–254 (1981)
33. Basri, R., Roth, D., and Jacobs, D.: Clustering appearances of 3D objects. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR-98), 414–420 (1998)
34. Roweis, S.T. and Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)
35. Kato, T.: Perturbation Theory for Linear Operators, 2nd edition. Springer (1980)
36. Nadler, B. and Galun, M.: Fundamental limitations of spectral clustering. NIPS, **19** (2006)
37. Nadler, B.: Finite Sample Convergence Results for Principal Component Analysis: A Matrix Perturbation Approach, submitted.
38. Zhou, D., Bousquet, O.,Navin Lal, T., Weston J., and Scholkopf, B.: Learning with local and global consistency. NIPS, **16** (2004)
39. Kevrekidis, I.G., Gear, C.W., Hummer, G.: Equation-free: The computer-aided analysis of complex multiscale systems. AIChE J. **50** 1346–1355 (2004)