



Diffusion-based kernel methods on Euclidean metric measure spaces



Amit Bermanis^a, Guy Wolf^b, Amir Averbuch^{b,*}

^a Department of Applied Mathematics, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel

^b School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

ARTICLE INFO

Article history:

Received 25 September 2014

Received in revised form 23 July 2015

Accepted 25 July 2015

Available online 30 July 2015

Communicated by Spec. Issue Guest Editor

Keywords:

Diffusion maps

Kernel methods

Diffusion-based kernel

Spectral analysis

Measure-based information

Manifold learning

ABSTRACT

Diffusion-based kernel methods are commonly used for analyzing massive high dimensional datasets. These methods utilize a non-parametric approach to represent the data by using an affinity kernel that represents similarities, distances or correlations between data points. The kernel is based on a Markovian diffusion process, whose transition probabilities are determined by local distances between data points. Spectral analysis of this kernel provides a representation of the data, where Euclidean distances correspond to diffusion distances between data points. When the data lies on a low dimensional manifold, these diffusion distances encompass the geometry of the manifold. In this paper, we present a generalized approach for defining diffusion-based kernels by incorporating measure-based information, which represents the density or distribution of the data, together with its local distances. The generalized construction does not require an underlying manifold to provide a meaningful kernel interpretation but assumes a more relaxed assumption that the measure and its support are related to a locally low dimensional nature of the analyzed phenomena. This kernel is shown to satisfy the necessary spectral properties that are required in order to provide a low dimensional embedding of the data. The associated diffusion process is analyzed via its infinitesimal generator and the provided embedding is demonstrated in two geometric scenarios.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

The utilization of kernel methods is a common practice in a non-parametric data analysis of massive high dimensional datasets. Usually, a limited set of underlying factors generates the high dimensional observable parameters via non-linear mappings. The non-parametric nature of this analysis overcomes the redundancies of the observable parameters and uncovers their underlying factors. These methods extend the well known MDS [9,18] method. They are based on a construction of an affinity kernel that encapsulates the relations

* Corresponding author. Fax: +972 153 54 5694455.

E-mail address: amir@math.tau.ac.il (A. Averbuch).

(distances, similarities or correlations) between data points. Spectral analysis of this kernel provides an efficient representation of the data that simplifies its analysis.

The MDS method uses the eigenvectors of a Gram matrix, which contains the inner products between the data points in the analyzed dataset, to define a mapping of data points into an embedded space that preserves most of these inner products. This method is equivalent to PCA [17,16], which projects the data onto the span of the principal directions of the variance of the data. Both of these methods capture linear structures on the data. They separate between meaningful directions, which represent the distribution of the data, and noisy uncorrelated directions. The former ones are associated with significant eigenvalues (and eigenvectors) of the Gram matrix, while the latter ones are associated with small eigenvalues.

Kernel methods, such as Isomap [28], LLE [25] and Laplacian eigenmaps [1], Hessian eigenmaps [12] and local tangent space alignment [29,31], extend the MDS paradigm by considering locally linear structures in the data. These structures are assumed to form a low dimensional manifold that captures the dependencies between the observable parameters of the data. This is called the manifold assumption, and the data is assumed to be sampled from this manifold. The spectral embedding space in these methods preserves the geometry of the manifold, which incorporates the underlying factors of the data.

The diffusion maps (DM) method [6] is a popular kernel method that utilizes a stochastic diffusion process to analyze the data. It defines diffusion affinities via symmetric conjugation of a transition probability operator. These probabilities are based on local distances between the data points. The Euclidean distances in the embedded space represent the diffusion distances in the original space. When the data is sampled from a low dimensional manifold, the diffusion paths follow the manifold and the diffusion distances capture its geometry.

In this paper, we enhance the DM method by incorporating information about the distribution of the data, in addition to local distances on which DM is based. This distribution is expressed in term of a measure over the observable space. The measure (and its support) replace the manifold assumption. We assume that the measure quantifies the likelihood for the presence of data over the geometry of the space. This assumption is significantly less restrictive than the need to have a manifold present. In practice this measure can either be provided as an input (e.g., by a-priori knowledge or a statistical model), or deduced from a given training set (e.g., by a density estimator). The manifold assumption can be expressed in terms of the measure assumption by setting the measure to be concentrated around an underlying manifold or (in the extremely restrictive case), to be supported by the manifold. Therefore, the measure assumption is not only less restrictive than the manifold assumption but it also generalizes it.

Data sampling densities were considered (and modeled by density measures) in previous variations of the DM framework, such as [6,10,11]. However, such sampling densities are typically an artifact resulting from nonuniform sampling of the underlying geometry, and the analysis does not use them to directly model the geometry of the data. Indeed, the anisotropic kernel [6], for example, is specifically aimed to separate the sampling density from the manifold geometry by either fully or partly canceling its effects on the diffusion process via appropriate kernel normalization. An alternative approach, presented in [10,11], is to use the sampling densities to locally adjust the diffusion scales, which determine the sizes of local data patches over the underlying geometry. In both these cases, the used densities are estimated directly from the sampled data that is used to construct the kernel, using unnormalized version of the kernel itself. In order for this density estimation to be accurate, large amounts of data are required both for measuring and for representing the densities. Such amounts are indeed commonly available in many Big Data applications. However, as in most kernel method, the size of the DM kernel is quadratically related to the size of dataset. Thus, computational requirements limit the sampled dataset size that can be effectively used for its construction in applicative settings. Therefore, tying the density estimation process directly to the kernel construction may be impractical.

In the suggested construction, the used measure is separated from the distances and from the analyzed dataset. As mentioned before, this measure can either represent densities or some other distribution

model determined by prior knowledge of the analyzed geometry. When dealing with discrete data, this construction can utilize two different sample sets of different sizes: the kernel domain that determines the kernel size, and the measure domain with attached empirical measure values. Furthermore, from a theoretical point of view, this construction combines continuous measures with either discrete or continuous datasets.

Similar to the case of sampled datasets from the manifold, for practical applications both the kernel domain and the measure domain would have to be appropriately discretized via sampling. The measure domain already has a measure attached to it, and we can assume that as a discrete set it is sampled according to this distribution. Additionally, since we design the kernel not to be affected by the size of this set, we can indeed consider it to be big enough to appropriately represent this distribution. The kernel domain, on the other hand, is assumed to be uniformly distributed, for simplicity, and therefore no sampling density is considered for it. This assumption is similar to the uniformity assumption in [1], and it can be achieved by a uniform grid construction, such as [3], or applying a whitening procedure on the analyzed data. The effects of nonuniform kernel domains will be considered in future work together with a related construction of adaptive nonuniform grids over such domains.

Alternative utilizations of two sets of data points together with DM were considered in two related approaches. First, one can use a small training set, which is either sampled from the entire data, or selected using dictionary construction (e.g. [13,26]), to alleviate the aforementioned computational limitations of DM. In this approach, the DM kernel and embedding are computed over the training set, and then extended, via interpolation, to the rest of the data, using out-of-sample methods such as [7,2,24]. However, the interpolation of the embedding does not affect the direct computation of the kernel (or the resulting DM), which only considers the training set and not the entire data. Therefore, the resulting analysis would still suffer from the same discussed limitations as the classical DM analysis.

Another approach, which is presented in [4,15], is to analyze the data by considering their relations to a given reference set, which can either be part of the input data, or designed for specific applications. Therefore, instead of representing pairwise similarities within the data, the kernel in these cases takes an asymmetric form consisting of relations between data points and reference points. It might be possible to utilize this approach in our settings by considering the measure domain (with the related measure) as the reference set, since the resulting asymmetric kernel would contain relations between the kernel domain and the measure domain. However, the size of the asymmetric kernel in this case would depend (albeit linearly) on the size of the measure domain, as well as the kernel domain. Therefore, practical considerations might limit its scalability in cases (such as common Big Data applications) where big data sets are considered as the measure domain.

In our approach, on the other hand, the constructed kernel is symmetric and it represents pairwise measure-based relations within the kernel domain. Therefore, in applicative settings the size of the constructed kernel matrix does not depend on the measure domain, but only on the size of the kernel domain. In continuous settings, it is an unweighted integral operator over the kernel domain, unlike the classic density-weighted DM one, since, as discussed above, we associate a uniform distribution with the kernel domain. Similar to DM and its variations, the suggested embedding is obtained by spectral analysis of a normalized version of the constructed kernel. This normalized kernel operator is shown to be a symmetric conjugate of a diffusion operator, and the properties of the corresponding diffusion process are analyzed and related to the measure that is used to model the properties of the data and the analyzed phenomena.

The structure of this paper is as follows. Section 2 describes the problem setup. A brief description of the DM method is presented in Section 2.1. Then, in Section 3, the measure-based kernel is formulated. Its spectral properties are presented in Section 3.1 and its infinitesimal generator is analyzed in Section 3.2. Finally, two geometric examples that demonstrate the proposed method are presented in Section 4.

2. Problem setup

Let $\Omega \subseteq \mathbb{R}^n$, for some natural n , be a compact metric space with the Euclidean distance metric. The integration notation $\int \cdot dy$ in this paper will refer to the Lebesgue integral $\int_{\Omega} \cdot dy$ over the subspace Ω , instead of the whole space $\mathbb{R}^n$. Let μ be a probability measure defined on Ω and let $q(x)$ be the density function of μ , i.e., $d\mu(x) = q(x)dx$. The measure μ and the density q are assumed to represent the distribution of data in Ω . Furthermore, we assume that q is sufficiently smooth¹ and it has a compact support $\text{supp}(q) \subset \Omega$, which is approximately locally low-dimensional. In the terminology presented in Section 1, the set Ω with a related uniform distribution will serve as the kernel domain, while $\text{supp}(q)$ with the related measure μ will be our measure domain.

We aim to combine the distance metric of Ω and the measure μ to define a kernel function $k(x, y)$, $x, y \in \Omega$, which represents the affinities between data points in Ω . Then, these affinities can be used to construct a diffusion map, as described in Section 2.1, and utilize it to embed the data into a low-dimensional representation that considers both proximities and distributions of the data points.

2.1. Diffusion maps

The diffusion maps (DM) framework utilizes a set of affinities to define a Markovian (random-walk) diffusion process over the analyzed data [6]. The spectral properties of this process are then used to obtain a representation of the data, where diffusion distances are expressed as Euclidean distances. The achieved representation reveals the underlying patterns of the data such as clusters and differences between normal and abnormal regions.

Technically, DM is based on an affinity kernel k and the associated integral operator that is defined as $Kf(x) = \int k(x, y)f(y)dy$, $x \in \Omega$, for any function $f \in L^2(\Omega)$. The affinity kernel k is normalized by a set of degrees $\nu(x) \triangleq \int k(x, y)dy$, $x \in \Omega$, to obtain the transition probabilities $p(x, y) \triangleq k(x, y)/\nu(x)$, from $x \in \Omega$ to $y \in \Omega$, of the Markovian diffusion process. Under mild conditions on the kernel k , the resulting transition probability operator has a discrete decaying spectrum of eigenvalues $1 = \lambda_0 \geq |\lambda_1| \geq |\lambda_2| \geq \dots$, which are used together with their corresponding eigenvectors $\vec{1} = \phi_0, \phi_1, \phi_2, \dots$ to achieve the diffusion map of the data.

Each data point $x \in \Omega$ is embedded by this diffusion map to the diffusion coordinates $(\lambda_1\phi_1(x), \dots, \lambda_\delta\phi_\delta(x))$, where the exact value of δ depends on the spectrum of the transition probabilities operator P , defined by $Pf(x) \triangleq \int p(x, y)f(y)dy$. The relation between the diffusion distance metric $\|p(x, \cdot) - p(y, \cdot)\|_{L^2(\Omega, \nu^{-1})}$ and the Euclidean distances in the embedded space, is a result of the spectral theorem [6,20]. Since the embedding is based on spectral analysis of the diffusion transition operator, it is usually comfortable to work with its symmetric conjugate $Af(x) \triangleq \int a(x, y)f(y)dy$, where $a(x, y) \triangleq \nu(x)^{1/2}p(x, y)\nu(y)^{-1/2} = k(x, y)/\sqrt{\nu(x)\nu(y)}$. This symmetric conjugate is called the diffusion affinity kernel, and the values $a(x, y)$, $x, y \in \Omega$, are the diffusion affinities of the data.

Usually, the Gaussian affinities $k_\varepsilon(x, y) = \exp(-\|x - y\|^2/2\varepsilon)$, for some suitable $\varepsilon > 0$, are used for the construction of the diffusion map. Here, $\|\cdot\|$ denotes the Euclidean metric in $\mathbb{R}^n$. When the data in Ω lies on a low dimensional manifold, its tangent spaces can be utilized to express the infinitesimal generator of the diffusion affinity kernel A in terms of the Laplacian operators on the manifold.

In this paper, we do not assume any underlying manifold structure. Instead, we assume we have access to a measure that represents the locally low dimensional distribution of the analyzed data. This measure can be supported by a low-dimensional manifold, but it can also represent non-manifold structures that have no tangent spaces. Another benefit of using a smooth measure instead of a strict underlying structure is that it can gradually dissipate, thus accounting for possible noise that results in data points being spread

¹ Specifically, we assume $q \in C^4(\Omega)$.

around an underlying structure, instead of strictly lying on that structure. The standard DM method, which is based on the Gaussian kernel, is unsuitable for this case since it only utilizes distances and does not inherently consider the measure μ , which is used in our case to model the geometry of data, and should not be confused with the unknown sampling density in [6], as explained in Section 1. In this paper, we will present an enhanced kernel that incorporates the measure information together with distance information to define affinities and utilize them to obtain the DM representation of the data in Ω .

3. Measure-based diffusion and affinity kernels

In this section, we define and analyze an affinity kernel that is based on the distances in Ω and on the measure μ . We use this kernel together with the DM method, which was briefly described in Section 2.1, to obtain a measure-based diffusion affinity kernel and its resulting diffusion map. In Section 3.1, we explore the spectral properties of the associated integral operator, which are crucial for the spectral analysis that provides the embedded diffusion coordinates. Then, in Section 3.2, we show the relations between the infinitesimal generator of the resulting diffusion operator and the Laplacian operator on Ω and the measure μ .

In order to define the desired kernel, we first define the function $g_\varepsilon : \mathbb{R} \rightarrow \mathbb{R}$,

$$g_\varepsilon(t) \triangleq e^{-t^2/\varepsilon}. \quad (3.1)$$

Notice that the Gaussian kernel, which is usually used in the DM method, can be defined as

$$k_\varepsilon(x, y) \triangleq g_{2\varepsilon}(\|x - y\|),$$

and this definition will be used in the rest of the paper. Definition 3.1 uses the function g_ε to define an alternative kernel that incorporates both local distance information, as the Gaussian kernel does, and measure information, which the Gaussian kernel lacks.

Definition 3.1 (*Measure-based Gaussian correlation kernel*). The Measure-based Gaussian Correlation (MGC) affinity function $\tilde{k}_\varepsilon : \Omega \times \Omega \rightarrow \mathbb{R}$ is defined as

$$\tilde{k}_\varepsilon(x, y) \triangleq \int g_\varepsilon(\|x - r\|) \cdot g_\varepsilon(\|y - r\|) d\mu(r).$$

The MGC integral operator is defined by this function as $\tilde{K}_\varepsilon f(x) = \int \tilde{k}_\varepsilon(x, y) f(y) dy$ for every function $f \in L^2(\Omega)$ and data point $x \in \Omega$.

Let us consider a discrete setup in which both Ω and $\text{supp}(q)$ are finite, i.e. $\Omega = \{x_1, \dots, x_n\}$ and q is a discrete measure with $\text{supp}(q) = \{r_1, \dots, r_s\}$. In this case, instead of the continuous kernel $\tilde{k}_\varepsilon$, we get an $n \times n$ kernel matrix $\tilde{K}_\varepsilon = G_\varepsilon Q G_\varepsilon^T$, where G is an $n \times s$ matrix whose (i, j) -th entry is $(G_\varepsilon)_{ij} \triangleq g_\varepsilon(\omega_i, r_j)$, $i = 1, \dots, n$, $j = 1, \dots, s$, and Q is a diagonal matrix, whose j -th diagonal element is $q(r_j)$, $j = 1, \dots, s$. The matrix G encapsulates the relations between the elements of the kernel domain and the elements of the measure domain. Thus, we get $\tilde{K}_\varepsilon = B_\varepsilon B_\varepsilon^T$, where $B_\varepsilon \triangleq G_\varepsilon \sqrt{Q}$ is $n \times s$ matrix.

Technically, this kernel form has been considered in several related works such as [30,14,21,19,15,27] and references therein. These works consider the relations between the analyzed dataset and a reference set, which typically is significantly smaller than the dataset. Then, the relations between the reference set and the dataset are used to construct a matrix B that is similar to B_ε . This matrix considers a big kernel $B^T B$ that captures the relations between data points via reference points. However, due to the big size of this kernel, which depends quadratically on the size of the dataset, it is usually not directly considered but examines the alternative kernel $B B^T$, which captures the relations between reference points via the dataset and extends the results to the big kernel. This smaller kernel is similar in nature to the MGC kernel from

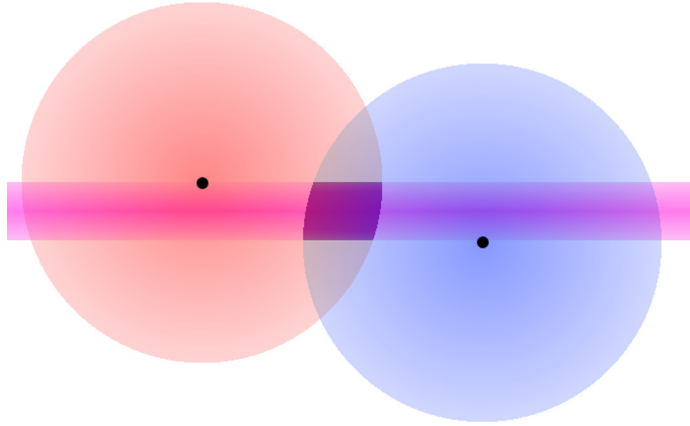


Fig. 3.1. An illustration of the MGC affinities. It shows a measure that surrounds a straight line (marked in magenta in the web version of the paper and right circle in the print version), and the Gaussians around two examined data points (marked in red and blue in the web version and left circle in the print version). The MGC affinity is based on the intersection (marked in dark purple in the web version and middle square in the print version) between the supports of these three functions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Definition 3.1, where the dataset serves as the measure domain and the reference set serves as the kernel domain. However, unlike these related works, in our case, we directly analyze the MGC kernel and the resulting diffusion geometry without considering a dual bigger kernel. The presented analysis in this paper establishes the significant information captured by this kernel and its importance in its own right.

The MGC affinity $\tilde{k}_\varepsilon(x, y)$, $x, y \in \Omega$, from **Definition 3.1**, is in fact the inner product in $L^2(\Omega, \mu)$ between two Gaussians of width ε that are centered at x and y , respectively. This affinity is based on the correlation, which also takes into consideration the measure μ , between the described Gaussians around at the examined data points as illustrated in **Fig. 3.1**. The numerically significant positions of r in this correlation must be close enough to x and to y (based on their Gaussians of radius ε), but they must also be in an area with a high enough concentration of the measure μ . Notice that the measure information is considered and incorporated in the affinity definitions. It is not required any more in the application of the kernel operator $\tilde{K}$ to functions over Ω . An alternative formulation of the MGC affinities is presented in **Proposition 3.1**.

Proposition 3.1. *The MGC affinities from **Definition 3.1** can also be expressed by*

$$\tilde{k}_\varepsilon(x, y) = k_\varepsilon(x, y) \cdot \int_{g_{\varepsilon/2}} \left(\left\| \frac{x+y}{2} - r \right\| \right) d\mu(r). \quad (3.2)$$

Proof. Using the identity $\|x - r\|^2 + \|y - r\|^2 = \frac{1}{2} \|x - y\|^2 + 2 \left\| \frac{x+y}{2} - r \right\|^2$, we get $g_\varepsilon(\|x - r\|)g_\varepsilon(\|y - r\|) = g_{2\varepsilon}(\|x - y\|)g_{\varepsilon/2}(\left\| \frac{x+y}{2} - r \right\|)$, which satisfies Eq. (3.2). $\square$

Proposition 3.1 shows the relation between the MGC kernel and the Gaussian kernel. While the Gaussian affinity only considers the distances between the examined data points, the MGC affinity also considers the region in which this distance is measured by using a Gaussian around the midpoint between them. This midpoint represents the direct path that determines the distance between the two data points. For a given distance between two data points, the MGC affinity increases when its path lies in an area with a high concentration of the measure μ , and decreases when it lies in an area with a low concentration of μ . If the measure μ is uniform over Ω , then the MGC kernel becomes the same as the Gaussian kernel up to a constant term that depends only on ε and can be easily normalized.

Consider the case of uniform distribution

$$q(x) = \begin{cases} 1 & x \in \Omega' \\ 0 & \text{otherwise} \end{cases},$$

where $\Omega' \subset \Omega$ is an open and connected set of unit volume, i.e., $\text{vol}(\Omega') = \int_{\Omega'} dr = 1$. In this case, for every $x, y \in \Omega$, the measure term of the MGC affinities according to [Proposition 3.1](#) becomes

$$\int g_{\varepsilon/2} \left(\left\| \frac{x+y}{2} - r \right\| \right) d\mu(r) = \int_{\Omega'} g_{\frac{\varepsilon}{2}} \left(\left\| \frac{x+y}{2} - r \right\| \right) dr,$$

thus it does not represent any meaningful information about the data points x and y . Indeed, whenever the midpoint of x and y is far from the boundary of Ω' (with respect to ε)

$$\int_{\Omega'} g_{\frac{\varepsilon}{2}} \left(\left\| \frac{x+y}{2} - r \right\| \right) dr \approx \left(\frac{\pi\varepsilon}{2} \right)^{n/2}.$$

Therefore, we can normalize the MGC affinities $\tilde{k}_\varepsilon$ to get *the normalized MGC affinity*

$$\hat{k}_\varepsilon(x, y) \triangleq \varepsilon^{-n/2} \tilde{k}_\varepsilon(x, y) \quad (3.3)$$

that converges to the Gaussian affinity k_ε when the measure is uniform, but incorporates the measure in the affinity when it is not uniform. The MGC affinity $\tilde{k}_\varepsilon$ and its normalized version $\hat{k}_\varepsilon$ only differ by a normalization term, thus they can be used interchangeably and the achieved results are equivalently valid for both of them.

The dual representation of the MGC kernel in [Definition 3.1](#) and [Proposition 3.1](#) can be used to detect and consider several common patterns in data analysis directly from the initial construction of the kernel. [Fig. 3.2\(a\)](#) uses the formulation in [Definition 3.1](#) to illustrate a case when the data is concentrated in areas around a curve with significant curvatures. In this case, the affinity will be more affected by the distances over the path that follows the “noisy” curve and not by the directions that follow sparse areas and bypass the curve. [Fig. 3.2\(b\)](#) uses the formulation in [Proposition 3.1](#) to illustrate the affinities when the data is concentrated in two distinct clusters. In this case, we can see that the affinity between data points from different clusters is significantly reduced due to the measure even if they are relatively close.

Notice that in both illustrated cases, the density around the examined data points is similar, and the important information comes from considering the densities in the areas between them. This emphasizes a significant difference between the MGC kernel, the anisotropic kernel in [\[20\]](#) and the adaptive kernel from [\[11\]](#). The latter two approximate the densities around the compared data points and use these densities to normalize or adjust the affinity between them. However, when these data points lie in similarly significant densities, these adjustments do not take into account the areas between them. In practice, when dealing with finite sampled datasets, the MGC kernel does not require knowledge of the densities (or measure values) at the compared data points (x and y in [Definition 3.1](#)), which can be sampled independently from the inner integrand values (r in [Definition 3.1](#)), for which the densities are required. In fact, we can use two different sets: the kernel domain and the measure domain. The utilization of these two sets of samples will be demonstrated in [Section 4](#) together with additional examples.

[Section 3.1](#) shows that the presented MGC affinity kernel satisfies the spectral properties that are required (and assumed) in [\[6,20\]](#) for its utilization with the DM framework. These properties enable us to define a diffusion process that is based on the MGC affinities. Then, the resulting diffusion map is used to embed the data in a way that considers the distances and the measure distribution. [Section 3.2](#) analyzes the properties of the resulting diffusion process by examining the infinitesimal generator of its transition probabilities and relating it to the infinitesimal generator in [\[6\]](#).

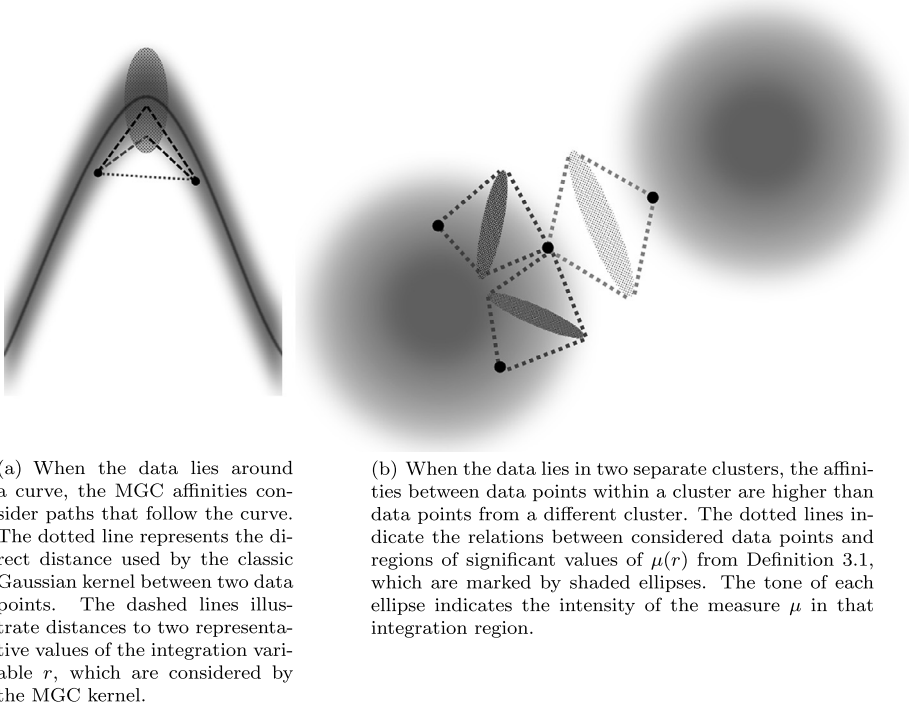


Fig. 3.2. An illustration of the MGC affinities in two common data analysis scenarios. For every pair of compared data points, the significant values of the integration variable r , from Definition 3.1 or Proposition 3.1, are marked by shaded ellipses.

3.1. Spectral properties

The DM embedding is based on spectral analysis of a normalized version of the used affinity kernel. Therefore, in order to use the MGC kernel with the DM analysis framework, the spectral properties of the associated integral operator have to be established first. In this section, we show that this kernel satisfies the assumptions (or conditions) in [6], thus, the achieved DM results are applicable when the MGC kernel is utilized to provide the affinities of the data.

We define the symmetric and positive kernel $\tilde{a}_\varepsilon : \Omega \times \Omega \rightarrow \mathbb{R}$ as

$$\tilde{a}_\varepsilon(x, y) \triangleq \frac{\tilde{k}_\varepsilon(x, y)}{\sqrt{\nu_\varepsilon(x)\nu_\varepsilon(y)}}, \quad (3.4)$$

where

$$\nu_\varepsilon(x) = \int \tilde{k}_\varepsilon(x, y) dy. \quad (3.5)$$

The normalization values $\nu_\varepsilon(x)$, $x \in \Omega$, are referred to as the diffusion degrees of the data. It should be noticed that since Ω is a compact set, the minimal value of $\nu_\varepsilon(x)$ retains its minimal value on Ω , and this value is positive. The associated integral operator is

$$\tilde{A}_\varepsilon f(x) \triangleq \int \tilde{a}_\varepsilon(x, y) f(y) dy. \quad (3.6)$$

This operator consists of the diffusion affinities of the data, when the diffusion is based on the MGC kernel. We will refer to it as the MGC diffusion affinities kernel. The operator $\tilde{A}_\varepsilon$ is the symmetric conjugate of a stochastic operator that consists of the transition probabilities of the underlying diffusion process as

was explained for the general DM setup in Section 2.1. Its symmetry eases the investigation of its spectral properties, which are (up to conjugacy) the properties of the conjugate stochastic one. Proposition 3.2 shows that $\tilde{A}_\varepsilon$ is a Hilbert–Schmidt operator.

Proposition 3.2. *The diffusion affinity operator $\tilde{A}_\varepsilon$ is a Hilbert–Schmidt operator from $L^2(\Omega)$ into itself where its norm is $\|\tilde{A}_\varepsilon\|_{L^2(\Omega)} = 1$. It is achieved by the square root of the stationary distribution of the underlying diffusion process.*

Corollary 3.3 is a direct consequence of Proposition 3.2. It essentially means that the spectral analysis of $\tilde{A}_\varepsilon$ results in a small number of significant eigenvalues (and eigenvectors). Therefore, this operator enables the utilization of the DM framework for dimensionality reduction based on the MGC affinities.

Corollary 3.3. *As a Hilbert–Schmidt operator, $\tilde{A}_\varepsilon$ is compact self-adjoint, therefore its spectrum is discrete, it decays to zero and it is bounded from above by 1.*

We can now prove Proposition 3.2.

Proof of Proposition 3.2. The kernel function $\tilde{k}_\varepsilon$ (see Definition 3.1) is positive and continuous on its support, therefore, $\nu_\varepsilon(x)$ satisfies the same properties. As a consequence, $\tilde{a}_\varepsilon(x, y)$ is a continuous function in the compact domain $\Omega \times \Omega$. Consequently, $\tilde{A}_\varepsilon$ is a Hilbert–Schmidt operator from $L^2(\Omega)$ into itself. Additionally,

$$\begin{aligned} \left| \int \tilde{k}_\varepsilon(x, y) \frac{f(y)}{\sqrt{\nu_\varepsilon(y)}} dy \right| &\leq \left(\int \tilde{k}_\varepsilon(x, y) dy \right)^{1/2} \left(\int \tilde{k}_\varepsilon(x, y) \frac{f^2(y)}{\nu_\varepsilon(y)} dy \right)^{1/2} \\ &= \sqrt{\nu_\varepsilon(x)} \left(\int \tilde{k}_\varepsilon(x, y) \frac{f^2(y)}{\nu_\varepsilon(y)} dy \right)^{1/2}, \end{aligned}$$

therefore,

$$\begin{aligned} |\langle \tilde{A}_\varepsilon f, f \rangle| &= \left| \iint \frac{\tilde{k}_\varepsilon(x, y)}{\sqrt{\nu_\varepsilon(x)\nu_\varepsilon(y)}} f(x)f(y) dx dy \right| \\ &\leq \int |f(x)| \left(\int \tilde{k}_\varepsilon(x, y) \frac{f^2(y)}{\nu_\varepsilon(y)} dy \right)^{1/2} dx \\ &\leq \|f\| \left[\iint \tilde{k}_\varepsilon(x, y) \frac{f^2(y)}{\nu_\varepsilon(y)} dy dx \right]^{1/2} \\ &= \|f\|^2, \end{aligned}$$

hence, $\|\tilde{A}_\varepsilon\|_{L^2(\Omega)} \leq 1$. Applying $\tilde{A}_\varepsilon$ to $\sqrt{\nu_\varepsilon(x)}$ yields

$$\tilde{A}_\varepsilon \left(\sqrt{\nu_\varepsilon(x)} \right) = \int \frac{\tilde{k}_\varepsilon(x, y)}{\sqrt{\nu_\varepsilon(x)}} dy = \sqrt{\nu_\varepsilon(x)}.$$

The stationary distribution is achieved by normalizing the degrees of the data by the volume $\int \nu(x) dx$. The last result remains valid even after normalization by this volume, thus the square root of the resulting stationary distribution is also an eigenvector of $\tilde{A}_\varepsilon$, associated with eigenvalue 1, as the proposition states. $\square$

Corollary 3.3 ensures that the DM can be utilized using the MGC kernel for dimensionality reduction. Furthermore, since the spectrum of $\tilde{A}_\varepsilon$ is bounded from above by 1, the diffusion process converges over time. **Proposition 3.4** shows that $\tilde{A}_\varepsilon$ is positive definite, therefore, the discrete spectrum of $\tilde{A}_\varepsilon$ lies in the interval $[0, 1]$.

Proposition 3.4. *The operator $\tilde{A}_\varepsilon$ is positive definite in $L^2(\Omega)$.*

Proof. For any $f \in L^2(\Omega)$ using **Definition 3.1** and Eq. (3.2)

$$\begin{aligned} \langle \tilde{A}_\varepsilon f, f \rangle &= \iint \tilde{k}_\varepsilon(x, y) \frac{f(y)}{\sqrt{\nu_\varepsilon(y)}} dy \frac{f(x)}{\sqrt{\nu_\varepsilon(x)}} dx \\ &= \iiint g_\varepsilon(\|x - r\|) g_\varepsilon(\|y - r\|) q(r) dr \frac{f(x)}{\sqrt{\nu_\varepsilon(x)}} \frac{f(y)}{\sqrt{\nu_\varepsilon(y)}} dx dy \\ &= \int \left\{ \int g_\varepsilon(\|x - r\|) \sqrt{q(r)} \frac{f(x)}{\sqrt{\nu_\varepsilon(x)}} dx \cdot \int g_\varepsilon(\|y - r\|) \sqrt{q(r)} \frac{f(y)}{\sqrt{\nu_\varepsilon(y)}} dy \right\} dr \\ &= \int \left(\int g_\varepsilon(\|x - r\|) \sqrt{q(r)} \frac{f(x)}{\sqrt{\nu_\varepsilon(x)}} dx \right)^2 dr \geq 0 \quad \square \end{aligned}$$

To conclude this section, we summarize the spectral properties of the MGC integral operator $\tilde{A}_\varepsilon$. The spectrum of this operator is discrete, positive, bounded from above by 1 and decays to zero. Therefore, the eigenvalues of $\tilde{A}_\varepsilon$ are denoted by $1 = \tilde{\lambda}_0 \geq \tilde{\lambda}_1 \geq \dots \geq 0$. These properties enable the utilization of the DM for dimensionality reduction, by using the MGC affinities kernel. More specifically, considering the eigensystem of $\tilde{A}_\varepsilon$, which satisfies $\tilde{A}_\varepsilon \tilde{\phi}_j = \tilde{\lambda}_j \tilde{\phi}_j$, $j = 0, 1, \dots$, the map² $x \mapsto (\tilde{\lambda}_1^t \tilde{\phi}_1(x), \dots, \tilde{\lambda}_\delta^t \tilde{\phi}_\delta(x))$ is well defined and converges as t tends to infinity.

3.2. Infinitesimal generator

The DM framework is based on Markovian diffusion process, which is defined and represented by a transition probability operator denoted by P_ε . The infinitesimal generator of this operator encompasses the nature of the diffusion process. In [6,20], it was shown that when the data is sampled from a low dimensional underlying manifold, the infinitesimal generator of P_ε has the form of *Laplacian+Potential*. In this section, we show a similar result, when using the MGC-based diffusion without requiring the underlying manifold assumption to hold.

The MGC affinity function $\tilde{k}_\varepsilon$ is symmetric and positive, i.e., $\tilde{k}_\varepsilon(x, y) > 0$ for any pair of data points $x, y \in \Omega$. To convert it to be a transition kernel of a Markov chain on Ω , we normalize it as follows:

$$\tilde{p}_\varepsilon(x, y) \triangleq \frac{\tilde{k}_\varepsilon(x, y)}{\nu_\varepsilon(x)}, \quad (3.7)$$

thus,

$$\int \tilde{p}_\varepsilon(x, y) dy = 1. \quad (3.8)$$

We define the corresponding stochastic operator

² The value of δ is determined by the numerical rank of the MGC operator, and it plays the same role here as in the original DM framework (see Section 2.1).

$$\tilde{P}_\varepsilon f(x) \triangleq \int \tilde{p}_\varepsilon(x, y) f(y) dy. \quad (3.9)$$

This operator is conjugate to $\tilde{A}$, defined in Eq. (3.6), as their kernels satisfy the conjugacy relation $\tilde{a}_\varepsilon(x, y) = \nu_\varepsilon^{1/2}(x) \tilde{p}_\varepsilon(x, y) \nu_\varepsilon^{-1/2}(y)$. Therefore, their spectral qualities are identical up to conjugacy. More specifically, their spectra are identical, and the eigenfunctions are conjugated, i.e., if $\psi_\varepsilon(x)$ is an eigenfunction of $\tilde{P}_\varepsilon$ corresponding to eigenvalue λ_ε , then $\nu_\varepsilon^{1/2}(x) \psi_\varepsilon(x)$ is an eigenfunction of $\tilde{A}_\varepsilon$, corresponding to the same eigenvalue. Similar relations between the diffusion affinities kernel $a(x, y)$ and the transition probabilities kernel $p(x, y)$ were already introduced in Section 2.1 as the DM building blocks.

The infinitesimal generator of the diffusion transition operator $\tilde{P}_\varepsilon$ is defined as

$$\mathcal{L} \triangleq \lim_{\varepsilon \rightarrow 0} \frac{I - \tilde{P}_\varepsilon}{\varepsilon}.$$

We use the notation $\Delta_\varepsilon = (I - \tilde{P}_\varepsilon)/\varepsilon$, thus the infinitesimal generator takes the form $\mathcal{L} = \Delta_{\varepsilon \rightarrow 0}$. Theorem 3.5 shows that the operator $\mathcal{L}$ takes the form *Laplacian+potential*, which is similar to the result shown in [20, Corollary 2]. The expression for $\mathcal{L}$, which Theorem 3.5 provides, characterizes the differential equation for diffusion processes [5,8], as discussed later in this section. Furthermore, examining this expression shows that the Laplacian term in fact captures information regarding the kernel domain Ω , while the potential term captures information regarding the measure $d\mu(x) = q(x)dx$.

Theorem 3.5. *If the density function q is in $C^4(\Omega)$, then the infinitesimal generator $\mathcal{L}$ of the MGC-based diffusion operator is*

$$\mathcal{L}f = -\frac{m_2}{m_0} \left(\Delta f + \left\langle \frac{\nabla q}{q}, \nabla f \right\rangle \right), \quad f \in C^4(\Omega),$$

where,

$$m_0 = \int g_1(\|x\|) dx,$$

$$m_2 = \int g_1(\|x\|) (x^{(j)})^2 dx,$$

and $x^{(j)}$ is the j -th component of x , for an arbitrary³ choice of $1 \leq j \leq n$.

The proof of Theorem 3.5 contains two parts. The first part, in Lemma 3.6, examines the application of the diffusion transition operator $\tilde{P}_\varepsilon$ to an arbitrary function. The second part, in Lemma 3.7, examines the asymptotic infinitesimal behavior of the operator Δ_ε , which results in the infinitesimal generator $\mathcal{L}$.

Lemma 3.6. *For any $x, y \in \Omega$ and for any positive ε*

$$\begin{aligned} \tilde{P}_\varepsilon f(x) &= (\varepsilon^n m_0^2 f(x) q(x) + m_0 m_2 \varepsilon^{n+1} (f(x) \Delta q(x) \\ &\quad + \nabla(q(x) \nabla f(x))) + O(\varepsilon^{n+2})) / \nu_\varepsilon(x). \end{aligned}$$

The proof of Lemma 3.6 is based on Taylor expansions of the function f and the density function q (of the measure). It is similar to the approach taken by [20,6], but instead of using tangential structures (of a manifold), we use measure based considerations. The complete proof is rather technical and it appears

³ Notice that the value of m_2 does not depend on the exact choice of j , since the Gaussian g_1 is isotropic.

in Appendix A.1. Lemma 3.7 uses the result in Lemma 3.6, to examine asymptotic behavior of the transition operator $\tilde{P}_\varepsilon$.

Lemma 3.7. *For any $x \in \Omega$,*

$$\lim_{\varepsilon \rightarrow 0} \frac{f(x) - \tilde{P}_\varepsilon f(x)}{\varepsilon} = -\frac{m_2 \nabla(q(x) \nabla f(x))}{m_0 q(x)}.$$

The proof of Lemma 3.7 relies on Lemma 3.6 and some technical limit calculations. The complete proof of this proposition appears in Appendix A.1. Theorem 3.5 is a direct result of Lemma 3.7. Indeed, $\nabla(q \nabla f) = \Delta f + \langle \nabla q, \nabla f \rangle$, which gives the expression for $\mathcal{L}$ in the theorem.

The meta-parameter ε , which determines the localization scale of the MGC kernel, can also be considered as a quantization scale of the diffusion time steps. Indeed, as $\varepsilon \rightarrow 0$ the discrete diffusion steps propagate to infinitesimally smaller local neighborhoods. Equivalently, these diffusion steps correspond to shorter time periods, until the process becomes continuous. Thus, one can consider $\mathcal{L}$ as a time derivative (with respect to ε) on the diffusion process. Therefore, by considering an appropriate potential $U(x)$, which is determined by the measure, the result from Theorem 3.5 provides the formulation

$$\mathcal{L}f \propto \Delta f - \langle \nabla U, \nabla f \rangle,$$

which is equivalent to the backward Fokker–Plank equation $\dot{f}(x, t) = -\nabla U(x) + \sqrt{2}\dot{\omega}(x, t)$, where $\omega(x, t)$ is a standard Brownian process, as the stochastic differential equation that governs the diffusion process. For further discussion and details on such formulations of diffusion processes we refer the reader to [20,22,23].

4. Geometric examples

In this section, we demonstrate by two examples the MGC kernel and the resulting diffusion map. The first example analyzes noisy data that is spread around a spiral curve. In this case, we compare the MGC kernel and its diffusion to the “classic” DM [6]. The second example presents a case when only the measure is given, and the analyzed data points are given by a uniform grid around the support of the measure. This case can occur, for example, when only statistical information about the distributions of the data is given, or when dealing with massive datasets where the analysis of individual data points is unfeasible. In this case, the original DM method from [6] cannot be applied at all since the distances of the uniform grid are meaningless. However, the MGC kernel is also based on measure information, therefore, it reveals the underlying geometry that is represented by this measure. This example demonstrates a case where in Theorem 3.5 only the potential term captures meaningful geometrical information (the Laplacian is only affected by the uniform grid), while the first example shows a case where both terms may be important.

Note: The figures in this section use three color maps. For reference, these color maps are presented in Fig. B.1 in Appendix B.

4.1. Noisy spiral curve

In this section, we compare the Gaussian-based DM embedding [6] with the embedding achieved by the MGC-based DM presented in this paper. We use a noisy spiral curve (see Fig. 4.1(a)) for the comparison. The dataset was produced by sampling 500 equally spaced points from the curve and then sampling 10 normally distributed data points around each of these curve points. The resulting data has 5000 data points that lie in areas around the curve, as shown in Fig. 4.1(a), where the curve is marked in red and the noisy data points are marked in blue. We used the same scale meta-parameter ε to the compared DM applications.

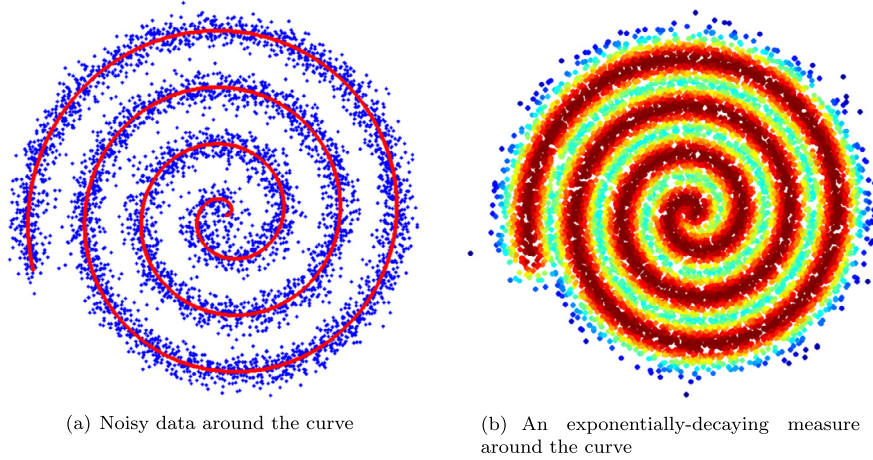


Fig. 4.1. A spiral curve with 5000 noisy data points concentrated around it, and 10^4 points that represent an exponentially-decaying measure around the curve. The colorscale color map from Fig. B.1(a) is used to represent the measure values. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

This meta-parameter was set to be sufficiently high to overcome the noise and to detect the high affinity between data points that originated from the same position (out of the 500 curve points) on the curve. Reasonable perturbations of this value do not significantly affect the presented results.

The application of the Gaussian-based DM is straight forward, as explained in [6]. The Gaussian kernel k is constructed and then normalized by the degrees to obtain the diffusion transition matrix P and the diffusion affinity matrix A . Spectral analysis of these matrices yields an embedding that is based on their most significant eigenvalues and eigenvectors.

The MGC kernel from Definition 3.1 requires to define a measure over the area where the data lies. Notice that the measure of the actual data points is not required. We can define a completely different set of points r from Definition 3.1 and then define their weights, which represent their measure values. We use two different measures for this definition. The first measure is based on 10^4 equally spaced data points from the curve and all the weights are set to one. This measure is essentially an indicator function of the spiral curve denoted by μ_c . The second measure is based on 10^4 points that are sampled around the curve by adding Gaussian noise to the data points that were used for defining μ_c . The weights of the point decay exponentially in relation to their distance from the curve. The resulting measure is denoted by μ_v and it is presented in Fig. 4.1(b) where the 10^4 measure points are colored according to their measure weights.

We use the notations $\tilde{K}_c$, $\tilde{P}_c$ and $\tilde{A}_c$ to denote the matrices that result from Definition 3.1, Eq. (3.9) and Eq. (3.6), respectively, with the measure μ_c . The notations $\tilde{K}_v$, $\tilde{P}_v$ and $\tilde{A}_v$ are used in a similar way for the measure μ_v . Notice that in both cases, even though the measure is based on 10^4 positions of the integration variable r (from Definition 3.1), the kernel and its normalized versions are of size 5000×5000 , since the data has only 5000 data points.

Fig. 4.2 compares the neighborhoods that are represented by the three kernels K , $\tilde{K}_c$ and $\tilde{K}_v$. We examine the neighborhoods of two data points on two different levels of the curve. In both cases, the Gaussian kernel captures inter-level affinities (i.e., it links different levels of the spiral) while both versions of the MGC kernel only capture relations in the same level of the spiral, thus, they are able to separate between these levels. In addition, the shape of the neighborhoods of the MGC kernels form ellipses whose major axes clearly follow the significant tangential directions of the curve. The Gaussian kernel, however, captures circular neighborhoods that do not express any information about the significant directions of the data. Since both K_c and K_v show similar neighborhoods and they indeed capture similar relations, we will only present from now on the comparison between the Gaussian-based diffusion and the MGC-based diffusion that is based on μ_c . Similar results are also achieved by using μ_v .

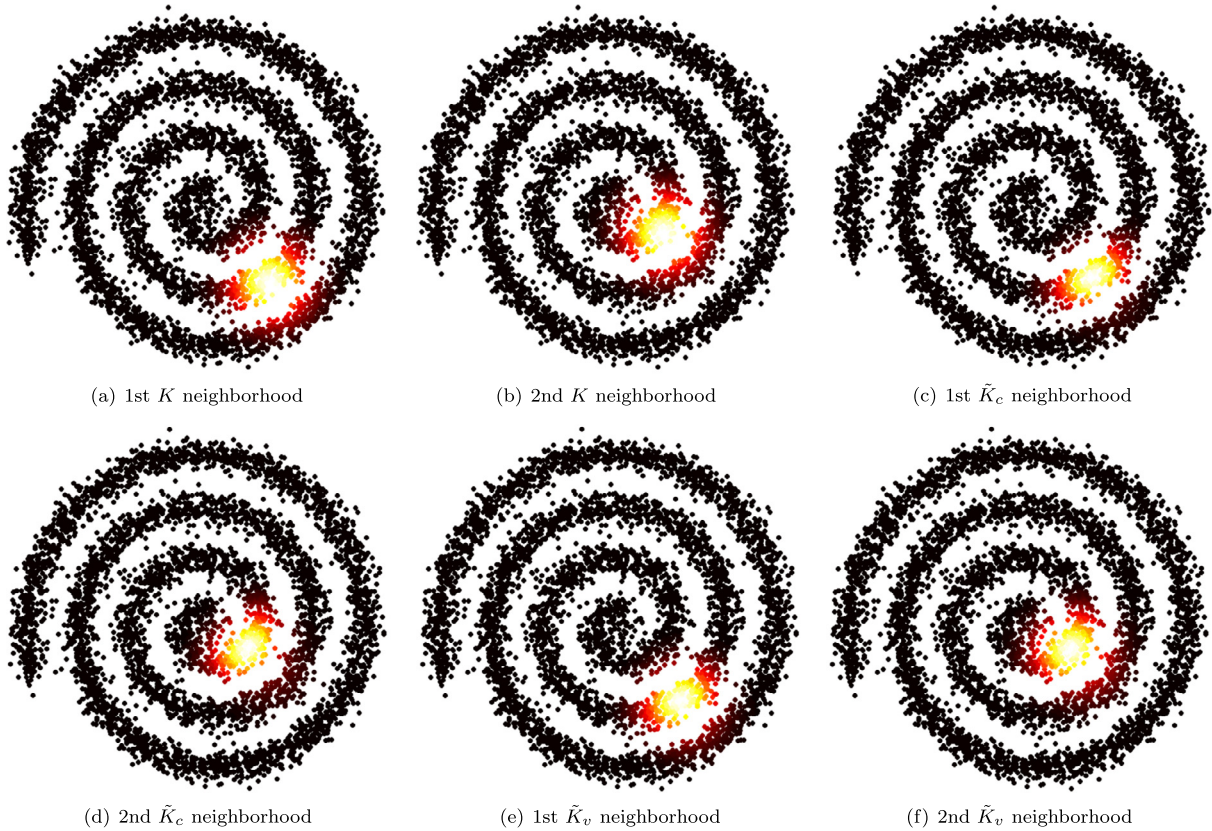


Fig. 4.2. Two neighborhoods from the Gaussian kernel (K) and the MGC kernels ($\tilde{K}_c$ and $\tilde{K}_v$) on the spiral curve, using the heat-map in Fig. B.1(b) to represent the kernel values.

The embedding, which is achieved by DM, is based on a diffusion process whose time steps are represented by powers of the diffusion transition matrix or the diffusion affinity matrix. The resulting Markov process has a stationary distribution when the time steps are taken to infinity. This stationary distribution reveals the concentrations and the underlying potential of the diffusion process. It is represented by the first eigenvector of the diffusion affinity matrix.⁴ Fig. 4.3 compares the stationary distributions of the Gaussian-based diffusion with the MGC-based diffusion as represented by the first eigenvector of the corresponding diffusion affinity matrix A or $\tilde{A}_c$. This comparison shows that the Gaussian-based diffusion considers the entire spiral as one pit of potential. At infinity, the diffusion is distributed over the entire region of the curve. The MGC-based diffusion, on the other hand, separates different levels of the spiral. At infinity, this diffusion is concentrated on the curve levels themselves and not on the areas between them.

Finally, we compare between the embedded spaces of the Gaussian-based DM and the MGC-based DM. Fig. 4.4 presents these embedded spaces based on the first two diffusion coordinates and Fig. 4.5 presents these spaces based on the first three diffusion coordinates (i.e., the two/three most significant eigenvectors of the diffusion transition operator). The comparison in Fig. 4.5 clearly shows that the MGC-based embedding results in a better separation between the spiral levels. Fig. 4.5 further establishes this observation by showing that, in fact, the Gaussian-based diffusion considers the whole noisy spiral as a two-dimensional disk. The MGC-based embedding, on the other hand, uses the third diffusion coordinate to completely separate the levels of the spiral by “stretching” it apart in the three-dimensional embedded space.

⁴ More accurately, the first eigenvector of the diffusion affinity is the square root of the stationary distribution, but it is sufficient to use these values for the purposes of these demonstrations.

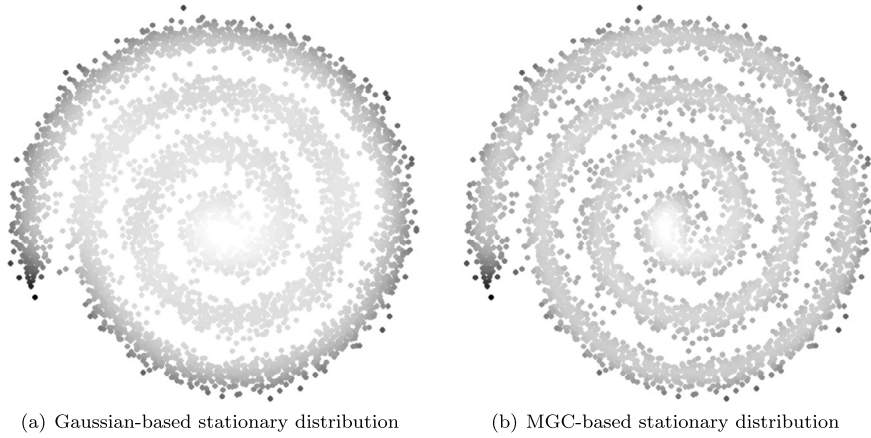


Fig. 4.3. The stationary distributions of: (a) the Gaussian-based diffusion process, and (b) the MGC-based diffusion process. Both use the gray scale color map from Fig. B.1(c) to represent the distribution values.

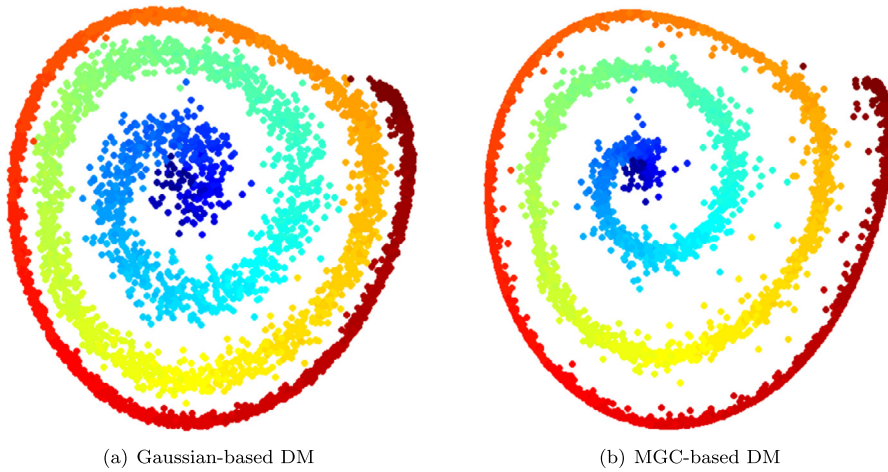


Fig. 4.4. The first two diffusion coordinates of the Gaussian-based and MGC-based DM embeddings.

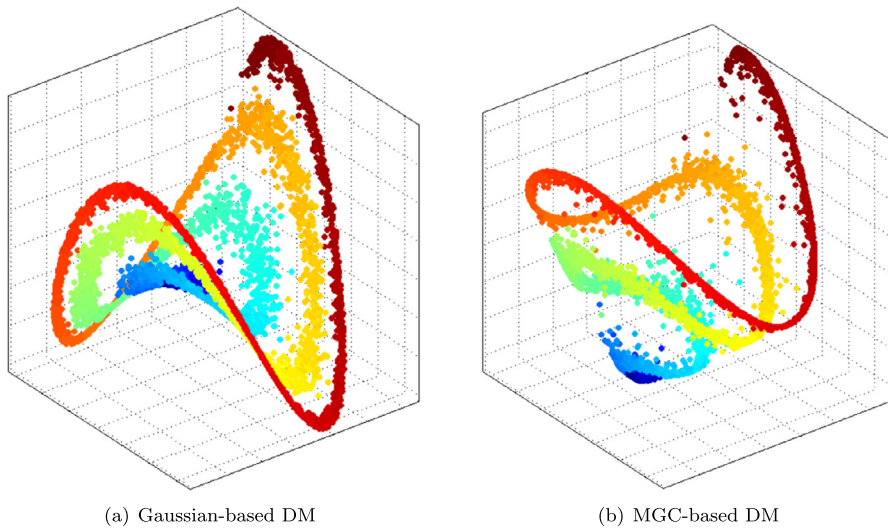


Fig. 4.5. The first three diffusion coordinates of the Gaussian-based and MGC-based DM embeddings.

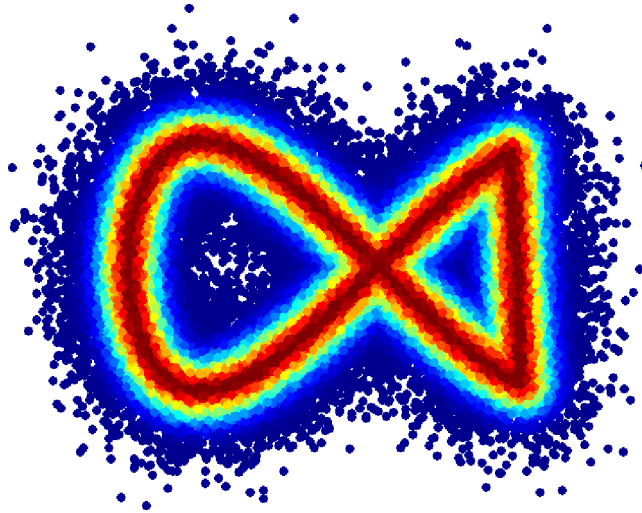


Fig. 4.6. Fish shape measure.

The superior results (e.g., separation between the spiral levels) of the MGC-based DM demonstrate its robustness to noise. The reason for this robustness is because the noise is part of the model on which the MGC construction is based. The Gaussian-based DM assumes that the data lies on (or it is sampled from) an underlying manifold, and any significant noise outside this manifold may violate this assumption. The MGC-based DM, on the other hand, already assumes variable concentrations and distributions of the data, which are represented by the measure and incorporated into the affinities. Therefore, this setting is more natural when dealing with data that is concentrated *around* an underlying manifold structure but does not necessarily lie on the manifold.

4.2. Uniform grid with a fish-shaped measure

In this section, we demonstrate a case when the Gaussian-based DM is inapplicable but the MGC-based DM can be applied for the analysis. Instead of using a discrete dataset of samples to represent the analyzed data, we use a measure, which holds the meaningful information about the analyzed phenomenon. This scenario can occur, for example, when dealing with massive datasets where it is unfeasible to analyze individual data points but one can obtain a density estimator over the observable space by using the massive number of samples. We will use a uniform grid of arbitrary size, which does not depend on the measure or its representative points, and utilize the MGC-based DM to analyze this grid in relation to the input measure.

We use a measure that is concentrated around a fish shape in two dimensions (see Fig. 4.6). It is represented by approximately 25,000 points. These points are sampled from areas around the support of the measure, and they are weighted according to their measure value. Fig. 4.6 shows the representative points and their measure-representing weights. In order to analyze the measure, we generate a 100×100 square grid in the bounding box of the support of the measure, and use the resulting 10,000 grid points as a dataset for the analysis. Since the grid is uniform, the distances between its grid points do not hold any meaningful information. Therefore, the Gaussian-based DM cannot be applied to analyze it. The MGC-based DM, on the other hand, can incorporate the measure information (based on the 25,000 representative data points) in the grid analysis. Thus, the resulting embedding will consider the meaningful information of the measure and not just the meaningless distances.

We use Definition 3.1 to construct the MGC kernel $\tilde{K}$ of the grid and the measure. The value of the meta-parameter ε was set to be sufficiently high for the uniform grid points to capture local measure information, and reasonable perturbations of this value do not significantly affect the presented results. The

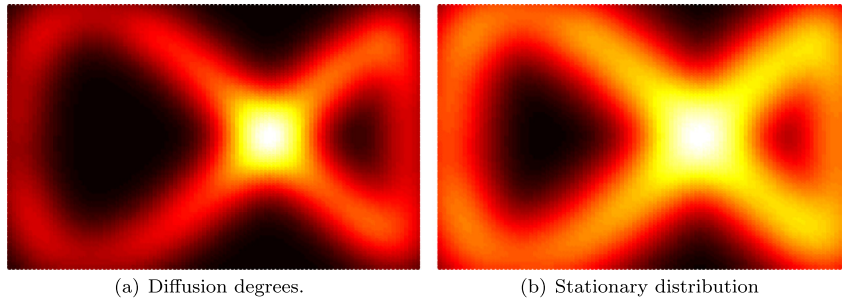


Fig. 4.7. Diffusion degrees and stationary distribution.

values of the integration variable r (in Definition 3.1) are taken from the 25,000 measure representatives, while the values of the compared points x and y (in Definition 3.1) are taken from the 10,000 grid points. The resulting kernel size is $10^4 \times 10^4$, and it does not depend on the number of measure representatives. Therefore, we can use an arbitrarily large number of points for representing the measure without affecting the MGC kernel size, which is only determined by the grid size.

In order to apply the DM scheme to the MGC kernel $\tilde{K}$, we normalize it to obtain the transition matrix $\tilde{P}$ (see Eq. (3.9)) and the diffusion affinity $\tilde{A}$ (see Eq. (3.6)). The normalization values of the kernel are the degrees of the data points in a graph that is represented by $\tilde{K}$ as its weighted adjacency matrix. These degrees measure the centrality of each data point in this graph and the resulting diffusion process. Fig. 4.7(a) shows the degrees of the grid data points. Even though the grid is uniform and its distances are meaningless, this figure shows that the data points that lie in concentrated areas of the measure, are more central than others. This property of the MGC-based construction is a result of the measure information being considered and incorporated in the MGC kernel.

Another property of the diffusion process is its stationary distribution. This distribution represents the underlying potential of the diffusion. It governs the concentrations of the diffusion process as it converges to an equilibrium. The stationary distribution is represented by the first eigenvector of the diffusion affinity $\tilde{A}$ and it is shown (for the grid data points) in Fig. 4.7(b). The result in this image is similar to the degrees shown in Fig. 4.7(a). The concentration areas of the diffusion process correspond to the concentration areas of the analyzed measure even though it used a dataset that is taken from a uniform grid whose distances are not related to the measure.

Finally, the embedded space of the MGC-based DM analysis is obtained by spectral analysis of the diffusion process. Fig. 4.8 shows the first (i.e., most significant) four diffusion coordinates of the embedded space. The first two coordinates clearly represent the head of the fish that was used to define the measure and the next two represent its tail. The relation between these two pairs of diffusion coordinates is better seen in Fig. 4.9, which shows that these two-dimensional subspaces are orthogonal. The MGC-based DM transformed the uniform grid by incorporating the measure information and detected the underlying fish-shaped structure.

The data in this example is spread uniformly over the observable space. The significant information comes from the fish-shaped measure. Furthermore, the fish shape, on which the measure is based, is not a manifold. The head and tail can be regarded as one-dimensional curves, although the tail has two sharp angles so it does not have tangent spaces in these points. However, the area that connects them has the shape of a cross and it is two dimensional. Unlike the Gaussian-based DM, the MGC-based DM is not based on an underlying manifold and it is able to incorporate the measure information and analyze the non-manifold fish-based measure by using the presented uniform grid-based dataset. The resulting embedded space in Figs. 4.8 and 4.9 shows that this analysis also detects the geometric properties of the underlying shape. The two one-dimensional parts of the fish are clearly detected and they are separated in two orthogonal two-dimensional subspaces in the embedded space.

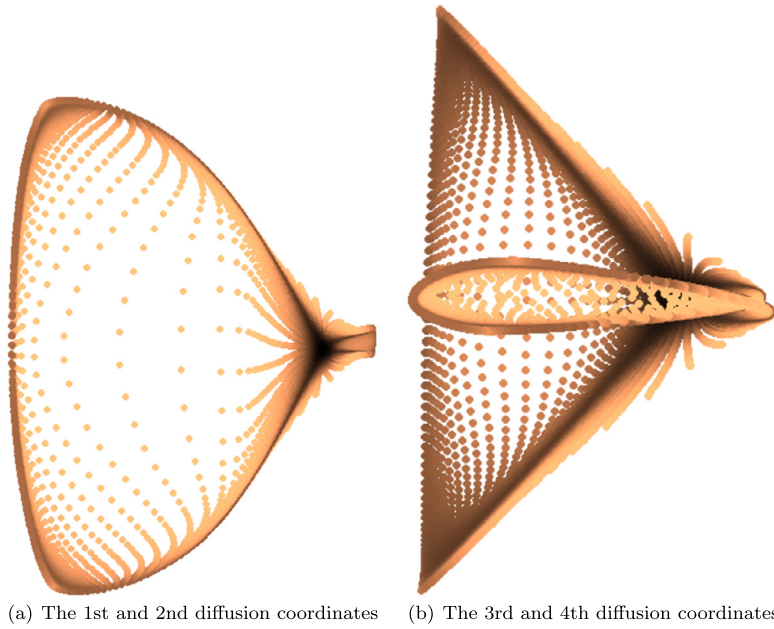


Fig. 4.8. The MGC-based DM embedding of the grid based on the first four MGC-based diffusion coordinates. It is presented in two pairs: 1st–2nd and 3rd–4th.

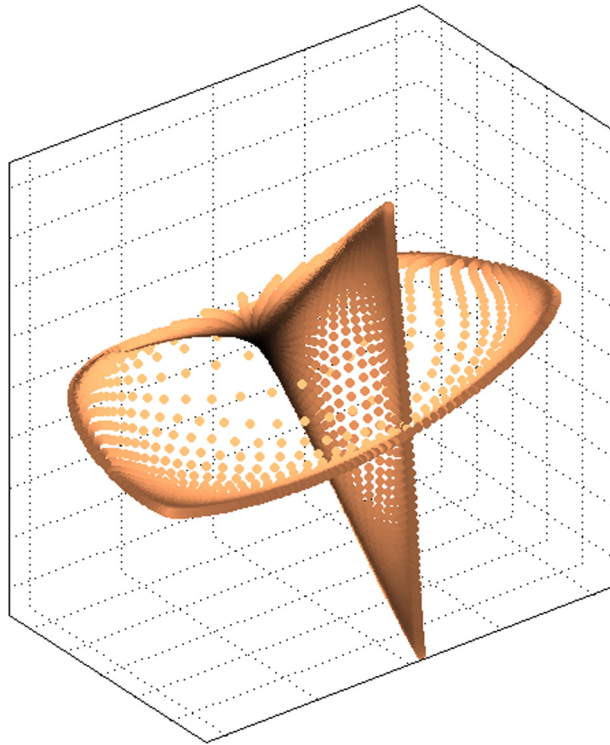


Fig. 4.9. The three-dimensional presentation of the embedded grid based on the second, third and fourth MGC-based diffusion coordinates.

5. Conclusion

We presented a generalized version of DM, which is based on the MGC kernel instead of the Gaussian kernel. We replaced the commonly-used manifold assumption in DM with a measure assumption. Namely,

we assume access to a measure that represents the locally low dimensional nature of the analyzed data, its distributions and its densities. The MGC kernel was presented and formulated in two equivalent forms that incorporate the measure-based information together with local distances between data points.

The spectral properties of the MGC-based construction of a diffusion map were explored and shown to be similar to the DM construction in [6]. These properties enable us to utilize the MGC-based DM for dimensionality reduction. Furthermore, we proved that the infinitesimal generator of the MGC-based diffusion process is similar to the diffusion process in [6], which is formed by a Laplacian operator and a diffusion potential. However, unlike the construction in [6], the MGC-based diffusion incorporates the measure information, which encompasses the distributions or densities of the data, in its transition probabilities. Therefore, this diffusion process inherently considers them.

We demonstrated the achieved embedding of the MGC-based DM in two scenarios. These examples demonstrated the robustness of the embedding to noise. This robustness is due to the noise being considered as part of the measure assumption while it violates the manifold assumption. Furthermore, we showed that the measure assumption can be used to capture non-manifold locally low dimensional structures with varying local dimensionality. Finally, since the MGC-based construction considers the measure and the data points separately, it is able to analyze a given measure distribution by using a uniform grid and deriving the represented underlying structure. This application cannot be achieved by the classic DM [6], which is based solely on local distances (up to possible nonuniform discretization artifacts) to recover the underlying geometry and does not consider a separately-provided measure. This aspect of the construction is further explored in [3], and the utilization of adaptive nonuniform grid constructions with the presented approach will be explored in future work.

Acknowledgments

This research was partially supported by the Israel Science Foundation (Grant No. 1041/10), the Israel Ministry of Science and Technology (Grants Nos. 3-9096, 3-10898) and Tel Aviv University Blavatnik Computer Science Funds. The second author was also supported by the Eshkol Fellowship from the Israeli Ministry of Science and Technology.

Appendix A. Technical proofs

A.1. Proof of Theorem 3.5

Theorem 3.5. *If the density function q is in $C^4(\Omega)$, then the infinitesimal generator $\mathcal{L}$ of the MGC-based diffusion operator is*

$$\mathcal{L}f = -\frac{m_2}{m_0} \left(\Delta f + \left\langle \frac{\nabla q}{q}, \nabla f \right\rangle \right), \quad f \in C^4(\Omega),$$

where,

$$m_0 = \int g_1(\|x\|) dx,$$

$$m_2 = \int g_1(\|x\|) (x^{(j)})^2 dx,$$

and $x^{(j)}$ is the j -th component of x , for an arbitrary⁵ choice of $1 \leq j \leq n$.

⁵ Notice that the value of m_2 does not depend on the exact choice of j , since the Gaussian g_1 is isotropic.

Since $f \in C^4(\Omega)$, it has a Taylor expansion of up to fourth order with bounded third order derivatives ($\text{supp}(\tilde{p}_\varepsilon)$ is compact). Hence, all continuous derivatives up to third order are bounded there. In the rest of the presentation, for a data point $x \in \Omega$, the notion $x^{(j)}$ means the j -th coordinate of x in a fixed coordinate system of Ω .

By change of variables, we get the following quantities for any $\varepsilon > 0$

$$\begin{aligned} \int g_\varepsilon(\|x\|) dx &= \varepsilon^{\frac{n}{2}} m_0, \\ \int g_\varepsilon(\|x\|) (x^{(j)})^2 dx &= \varepsilon^{\frac{n}{2}+1} m_2. \end{aligned} \quad (\text{A.1})$$

These integrals are used to prove [Lemma A.2](#), which provides an expression for the measure-related term of the MGC kernel (according to [Proposition 3.1](#)), in terms of m_0 , m_2 and the Laplacian of the density function q .

Lemma A.2. *If the density function q is in $C^4(\Omega)$ and its fourth derivatives are bounded, then, for any $x, y \in \Omega$ and any positive ε*

$$\begin{aligned} \int g_{\frac{\varepsilon}{2}} \left(\left\| \frac{x+y}{2} - r \right\| \right) q(r) dr &= m_0 \left(\frac{\varepsilon}{2} \right)^{\frac{n}{2}} q \left(\frac{x+y}{2} \right) \\ &\quad + \frac{1}{2} m_2 \left(\frac{\varepsilon}{2} \right)^{\frac{n}{2}+1} \Delta q \left(\frac{x+y}{2} \right) + O(\varepsilon^{\frac{n}{2}+2}). \end{aligned}$$

Proof. First we Taylor expand $q(r)$ around $\frac{x+y}{2}$, up to third order:

$$\begin{aligned} q(r) &= q \left(\frac{x+y}{2} \right) + \sum_{j=1}^n \partial_j q \left(\frac{x+y}{2} \right) \left(r^{(j)} - \frac{x^{(j)} + y^{(j)}}{2} \right) \\ &\quad + \frac{1}{2} \sum_{j=1}^n \partial_j^2 q \left(\frac{x+y}{2} \right) \left(r^{(j)} - \frac{x^{(j)} + y^{(j)}}{2} \right)^2 \\ &\quad + \sum_{i,j=1, i \neq j}^n \partial_i \partial_j q \left(\frac{x+y}{2} \right) \left(r^{(i)} - \frac{x^{(i)} + y^{(i)}}{2} \right) \left(r^{(j)} - \frac{x^{(j)} + y^{(j)}}{2} \right) \\ &\quad + \frac{1}{3!} \sum_{j=1}^n \partial_j^3 q \left(\frac{x+y}{2} \right) \left(r^{(j)} - \frac{x^{(j)} + y^{(j)}}{2} \right)^3 \\ &\quad + \frac{1}{2} \sum_{i,j=1, i \neq j}^n \partial_i \partial_j^2 q \left(\frac{x+y}{2} \right) \left(r^{(i)} - \frac{x^{(i)} + y^{(i)}}{2} \right) \left(r^{(j)} - \frac{x^{(j)} + y^{(j)}}{2} \right)^2 \\ &\quad + \sum_{i,j,k=1, i \neq j \neq k}^n \partial_i \partial_j \partial_k q \left(\frac{x+y}{2} \right) \left(r^{(i)} - \frac{x^{(i)} + y^{(i)}}{2} \right) \\ &\quad \times \left(r^{(j)} - \frac{x^{(j)} + y^{(j)}}{2} \right) \left(r^{(k)} - \frac{x^{(k)} + y^{(k)}}{2} \right) \\ &\quad + O \left(\left\| r - \frac{x+y}{2} \right\|^4 \right). \end{aligned}$$

When we integrate the above against $g_{\frac{\varepsilon}{2}}(\|\frac{x+y}{2} - r\|)$, all the first order, mixed second order and third order terms vanish. This is due to the fact that odd functions vanish when they are integrated against an even function in a symmetric domain (around $\frac{x+y}{2}$). Therefore,

$$\begin{aligned} \int g_{\frac{\varepsilon}{2}}\left(\left\|\frac{x+y}{2}-r\right\|\right) q(r) dr &= q\left(\frac{x+y}{2}\right) \int g_{\frac{\varepsilon}{2}}\left(\left\|\frac{x+y}{2}-r\right\|\right) dr \\ &\quad + \frac{1}{2} \sum_{j=1}^n \partial_j^2 q\left(\frac{x+y}{2}\right) \\ &\quad \times \int g_{\frac{\varepsilon}{2}}\left(\left\|\frac{x+y}{2}-r\right\|\right) \left(r^{(j)} - \frac{x^{(j)} + y^{(j)}}{2}\right)^2 dr \\ &\quad + \int g_{\frac{\varepsilon}{2}}\left(\left\|\frac{x+y}{2}-r\right\|\right) O\left(\left\|r - \frac{x+y}{2}\right\|^4\right) dr \end{aligned}$$

Thus, using Eq. (A.1), we get

$$\begin{aligned} \int g_{\frac{\varepsilon}{2}}\left(\left\|\frac{x+y}{2}-r\right\|\right) q(r) dr &= m_0\left(\frac{\varepsilon}{2}\right)^{\frac{n}{2}} q\left(\frac{x+y}{2}\right) \\ &\quad + \frac{1}{2} m_2\left(\frac{\varepsilon}{2}\right)^{\frac{n}{2}+1} \Delta q\left(\frac{x+y}{2}\right) + O(\varepsilon^{\frac{n}{2}+2}). \quad \square \end{aligned}$$

In the rest of this section, we assume that the assumptions of Lemma A.2 hold. Based on this lemma, Lemma 3.6 expresses the transition operator $\tilde{P}_\varepsilon$ in term of m_0 , m_2 and differential operators.

Lemma 3.6. *For any $x, y \in \Omega$ and for any positive ε*

$$\begin{aligned} \tilde{P}_\varepsilon f(x) &= (\varepsilon^n m_0^2 f(x) q(x) + m_0 m_2 \varepsilon^{n+1} (f(x) \Delta q(x) \\ &\quad + \nabla(q(x) \nabla f(x))) + O(\varepsilon^{n+2})) / \nu_\varepsilon(x). \end{aligned}$$

Proof. By combining Eq. (3.2) with Lemma A.2 we get

$$\begin{aligned} \int \tilde{k}_\varepsilon(x, y) f(y) dy &= m_0\left(\frac{\varepsilon}{2}\right)^{\frac{n}{2}} \int g_{2\varepsilon}(\|x - y\|) q\left(\frac{x+y}{2}\right) f(y) dy \\ &\quad + \frac{1}{2} m_2\left(\frac{\varepsilon}{2}\right)^{\frac{n}{2}+1} \int g_{2\varepsilon}(\|x - y\|) \Delta q\left(\frac{x+y}{2}\right) f(y) dy \\ &\quad + O(\varepsilon^{\frac{n}{2}+2}) \int g_{2\varepsilon}(\|x - y\|) f(y) dy. \end{aligned}$$

By multiplying the Taylor expansions of $q(\frac{x+y}{2})$ and $f(y)$ around x , we get

$$\begin{aligned} q\left(\frac{x+y}{2}\right) f(y) &= f(x) q(x) + \frac{1}{8} f(x) \sum_{j=1}^n \partial_j^2 q(x) \left(y^{(j)} - x^{(j)}\right)^2 \\ &\quad + \frac{1}{2} q(x) \sum_{j=1}^n \partial_j^2 f(x) \left(y^{(j)} - x^{(j)}\right)^2 \\ &\quad + \frac{1}{2} \sum_{j=1}^n \partial_j q(x) \left(y^{(j)} - x^{(j)}\right) \times \sum_{j=1}^n \partial_j f(x) \left(y^{(j)} - x^{(j)}\right) \\ &\quad + \text{odd function} + O(\|y - x\|^4) \end{aligned}$$

Notice that for $i \neq j$, the fourth addend is odd in respect to integration against $g_{2\varepsilon}(\|x - y\|)$, therefore, it vanishes. Thus, by using Eq. (A.1) we get

$$\begin{aligned}
 \int g_{2\varepsilon}(\|x - y\|) q\left(\frac{x + y}{2}\right) f(y) dy &= f(x) q(x) \int g_{2\varepsilon}(\|x - y\|) dy \\
 &+ \frac{1}{8} f(x) \sum_{j=1}^n \partial_j^2 q(x) \int g_{2\varepsilon}(\|x - y\|) \left(y^{(j)} - x^{(j)}\right)^2 dy \\
 &+ \frac{1}{2} q(x) \sum_{j=1}^n \partial_j^2 f(x) \int g_{2\varepsilon}(\|x - y\|) \left(y^{(j)} - x^{(j)}\right)^2 dy \\
 &+ \frac{1}{2} \sum_{j=1}^n \partial_j f(x) \partial_j q(x) \int g_{2\varepsilon}(\|x - y\|) \left(y^{(j)} - x^{(j)}\right)^2 dy + O(\varepsilon^{\frac{n}{2}+2}) \\
 &= m_0 (2\varepsilon)^{\frac{n}{2}} f(x) q(x) + \frac{1}{8} (2\varepsilon)^{\frac{n}{2}+1} m_2 \times (f(x) \Delta q(x)) \\
 &+ \frac{1}{2} (2\varepsilon)^{\frac{n}{2}+1} m_2 \times (q(x) \Delta f(x) + \nabla f(x) \cdot \nabla q(x)) + O(\varepsilon^{\frac{n}{2}+2}) \\
 &= m_0 (2\varepsilon)^{\frac{n}{2}} f(x) q(x) + \frac{1}{8} (2\varepsilon)^{\frac{n}{2}+1} m_2 \times (f(x) \Delta q(x)) \\
 &+ \frac{1}{2} (2\varepsilon)^{\frac{n}{2}+1} m_2 \times \nabla (q(x) \nabla f(x)) + O(\varepsilon^{\frac{n}{2}+2}). \tag{A.2}
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 \int g_{2\varepsilon}(\|x - y\|) \Delta q\left(\frac{x + y}{2}\right) f(y) dy &= m_0 (2\varepsilon)^{\frac{n}{2}} f(x) \Delta q(x) \\
 &+ \frac{1}{8} (2\varepsilon)^{\frac{n}{2}+1} m_2 \times (f(x) \Delta \Delta q(x)) + \frac{1}{2} (2\varepsilon)^{\frac{n}{2}+1} m_2 \\
 &\times \nabla (\Delta q(x) \nabla f(x)) + O(\varepsilon^{\frac{n}{2}+2}). \tag{A.3}
 \end{aligned}$$

In addition, since

$$\begin{aligned}
 f(y) &= f(x) + \frac{1}{2} \sum_{j=1}^n \partial_j^2 f(x) \left(y^{(j)} - x^{(j)}\right)^2 \\
 &+ \text{odd function} + O(\|y - x\|^4),
 \end{aligned}$$

we get

$$\begin{aligned}
 \int g_{2\varepsilon}(\|x - y\|) f(y) dy &= f(x) \int g_{2\varepsilon}(\|x - y\|) dy \\
 &+ \frac{1}{2} \sum_{j=1}^n \partial_j^2 f(x) \int g_{2\varepsilon}(\|x - y\|) \left(y^{(j)} - x^{(j)}\right)^2 dy \\
 &+ O(\varepsilon^{\frac{n}{2}+2}) \\
 &= (2\varepsilon)^{\frac{n}{2}} m_0 + \frac{1}{2} \varepsilon^{\frac{n}{2}+1} m_2 \Delta f(x) + O(\varepsilon^{\frac{n}{2}+2}). \tag{A.4}
 \end{aligned}$$

By summing up Eqs. (A.2)–(A.4), we get

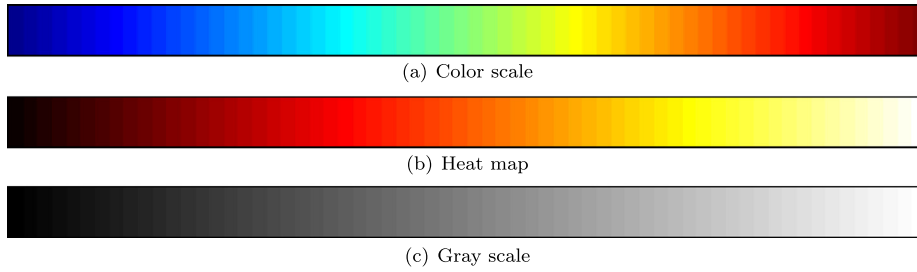


Fig. B.1. The color maps that are used in Section 4, where low values are on the left side and high values are on the right side of the map/scale.

$$\begin{aligned}
 \int \tilde{k}_\varepsilon(x, y) f(y) dy &= \varepsilon^n m_0^2 f(x) q(x) \\
 &\quad + m_0 m_2 \varepsilon^{n+1} (f(x) \Delta q(x) + \nabla(q(x) \nabla f(x))) \\
 &\quad + O(\varepsilon^{n+2}).
 \end{aligned} \tag{A.5}$$

Division by $\nu_\varepsilon(x)$ yields the result in the Lemma. $\square$

Lemma 3.7 provides an asymptotic expression for $\tilde{P}_\varepsilon$ in terms of m_0 , m_2 and differential operators. **Theorem 3.5** is a direct result of **Lemma 3.7** since $\nabla(q \nabla f) = q \Delta f + \langle \nabla q, \nabla f \rangle$ as stated in Section 3.2.

Lemma 3.7. For any $x \in \Omega$,

$$\lim_{\varepsilon \rightarrow 0} \frac{f(x) - \tilde{P}_\varepsilon f(x)}{\varepsilon} = -\frac{m_2 \nabla(q(x) \nabla f(x))}{m_0 q(x)}.$$

Proof. By substituting $f \equiv 1$ in Eq. (A.5), we have

$$\nu_\varepsilon(x) = \int \tilde{k}_\varepsilon(x, y) dy = \varepsilon^n m_0^2 q(x) + m_0 m_2 \varepsilon^{n+1} \Delta q(x) + O(\varepsilon^{n+2}).$$

Therefore,

$$\begin{aligned}
 \frac{1}{\varepsilon} [f(x) - \tilde{P}_\varepsilon f(x)] &= \frac{1}{\varepsilon} \left[f(x) - \frac{\int \tilde{k}_\varepsilon(x, y) f(y) dy}{\int \tilde{k}_\varepsilon(x, y) dy} \right] \\
 &= \frac{1}{\varepsilon} \left[f(x) - \frac{\varepsilon^n m_0^2 f(x) q(x) + m_0 m_2 \varepsilon^{n+1} (f(x) \Delta q(x) + \nabla(q(x) \nabla f(x))) + O(\varepsilon^{n+2})}{\varepsilon^n m_0^2 q(x) + m_0 m_2 \varepsilon^{n+1} \Delta q(x) + O(\varepsilon^{n+2})} \right] \\
 &= -\frac{m_0 m_2 \nabla(q(x) \nabla f(x)) + O(\varepsilon)}{m_0^2 q(x) + m_0 m_2 \varepsilon \Delta q(x) + O(\varepsilon^2)} \rightarrow -\frac{m_2 \nabla(q(x) \nabla f(x))}{m_0 q(x)} \quad \square
 \end{aligned}$$

Appendix B. Color maps

The color maps that are used in this paper, and specifically in Section 4, are shown in Fig. B.1.

References

- [1] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.* 15 (6) (2003) 1373–1396.
- [2] A. Bermanis, A. Averbuch, R.R. Coifman, Multiscale data sampling and function extension, *Appl. Comput. Harmon. Anal.* 34 (2013) 15–29.

- [3] A. Bermanis, M. Salhov, G. Wolf, A. Averbuch, Measure-based diffusion grid construction and high-dimensional data discretization, *Appl. Comput. Harmon. Anal.* 40 (2) (2016) 207–228.
- [4] R.R. Coifman, M.J. Hirn, Bi-stochastic kernels via asymmetric affinity functions, *Appl. Comput. Harmon. Anal.* 35 (1) (2013) 177–180.
- [5] R.R. Coifman, I.G. Kevrekidis, S. Lafon, M. Maggioni, B. Nadler, Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems, in: *Multiscale Modeling and Simulation*, 2008, pp. 842–864.
- [6] R.R. Coifman, S. Lafon, Diffusion maps, *Appl. Comput. Harmon. Anal.* 21 (1) (2006) 5–30.
- [7] R.R. Coifman, S. Lafon, Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions, *Appl. Comput. Harmon. Anal.* 21 (1) (2006) 31–52.
- [8] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps, *Proc. Natl. Acad. Sci. USA* 102 (21) (2005) 7426–7431.
- [9] T. Cox, M. Cox, *Multidimensional Scaling*, Chapman and Hall, London, UK, 1994.
- [10] G. David, Anomaly detection and classification via diffusion processes in hyper-networks, PhD thesis, School of Computer Science, Tel Aviv University, March 2009.
- [11] G. David, A. Averbuch, Hierarchical data organization, clustering and denoising via localized diffusion folders, *Appl. Comput. Harmon. Anal.* 33 (1) (2011) 1–23.
- [12] D.L. Donoho, C. Grimes, Hessian eigenmaps: new locally linear embedding techniques for high dimensional data, *Proc. Natl. Acad. Sci. USA* 100 (May 2003) 5591–5596.
- [13] Y. Engel, S. Mannor, R. Meir, The kernel recursive least-squares algorithm, *IEEE Trans. Signal Process.* 52 (8) (2004) 2275–2285.
- [14] A. Haddad, D. Kushnir, R.R. Coifman, Filtering via a reference set, Tech. rep. YALEU/DCS/TR-1441, Yale University, New Haven, CT, 2011.
- [15] A. Haddad, D. Kushnir, R.R. Coifman, Texture separation via a reference set, *Appl. Comput. Harmon. Anal.* 36 (2) (2014) 335–347.
- [16] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.* 24 (1933).
- [17] I.T. Jolliffe, *Principal Component Analysis*, Springer, New York, NY, 1986.
- [18] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* 29 (1964) 1–27.
- [19] D. Kushnir, A. Haddad, R.R. Coifman, Anisotropic diffusion on sub-manifolds with application to earth structure classification, *Appl. Comput. Harmon. Anal.* 32 (2) (2012) 280–294.
- [20] S. Lafon, Diffusion maps and geometric harmonics, PhD thesis, Yale University, May 2004.
- [21] W. Lian, R. Talmon, H. Zaveri, L. Carin, R.R. Coifman, Multivariate time-series analysis and diffusion maps, *Signal Process.* 116 (2015) 13–28.
- [22] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), *Advances in Neural Information Processing Systems*, vol. 18, MIT Press, Cambridge, MA, 2006, pp. 955–962.
- [23] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, *Appl. Comput. Harmon. Anal.* 21 (1) (2006) 113–127.
- [24] N. Rabin, R.R. Coifman, Heterogeneous datasets representation and learning using diffusion maps and Laplacian pyramids, in: *SDM*, 2012, pp. 189–199.
- [25] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* 290 (December 2000) 2323–2326.
- [26] M. Salhov, A. Bermanis, G. Wolf, A. Averbuch, Approximately-isometric diffusion maps, *Appl. Comput. Harmon. Anal.* 38 (3) (2015) 399–419.
- [27] R. Talmon, I. Cohen, S. Gannot, R.R. Coifman, Supervised graph-based processing for sequential transient interference suppression, *IEEE Trans. Audio, Speech Language Process.* 20 (9) (2012) 2528–2538.
- [28] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* 290 (5500) (2000) 2319–2323.
- [29] G. Yang, X. Xu, J. Zhang, Manifold alignment via local tangent space alignment, in: *International Conference on Computer Science and Software Engineering*, December 2008.
- [30] R. Zass, A. Shashua, A unifying approach to hard and probabilistic clustering, in: *Tenth IEEE International Conference on Computer Vision*, vol. 1, ICCV 2005, IEEE, 2005, pp. 294–301.
- [31] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, Technical report CSE-02-019, Department of Computer Science and Engineering, Pennsylvania State University, 2002.