University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Riki-Taavi Nurm

Diffusion Distance and Diffusion Maps: Theory and Applications

Mathematics
Bachelor's Thesis (9 EAP)

Supervisors:
Raul Vicente Zafra, Professor, University of Tartu
Johann Langemets, Associate Professor, University of Tartu

Tartu 2026

Contents

DIFFUSION DISTANCE AND DIFFUSION MAPS: THEORY AND APPLICATIONS

Bachelor's thesis

Riki-Taavi Nurm

Abstract

In computer vision and data analysis, we are often interested in studying the intrinsic geometry of a dataset independently of the ambient space in which it lives. After transforming the dataset into a Markov graph, we can define the diffusion distance between any two vertices based on how random walks diffuse between them. Using the eigenvalues and eigenvectors of the corresponding Markov transition matrix, we can define diffusion maps which embed the graph into a Euclidean space. The main aim of this thesis is to prove a theorem which states that the diffusion distance on the graph is equivalent, up to a scaling constant, to the Euclidean distance in the embedded space. We illustrate the theorem on synthetic and established datasets.

CERCS research specialisation: P170 Computer science, numerical analysis, systems, control.

Keywords: diffusion maps, diffusion distance, spectral theorem, Markov chains, dimensionality reduction.

DIFUSIOONKAUGUS JA DIFUSIOONIKAARDID: TEOORIA JA RAKENDUSED

Bakalaureusetöö

Riki-Taavi Nurm

Lühikokkuvõte

Arvutinägemises ja andmeanalüüsis tahame me tihti uurida andmete sisemist geomeetriat sõltumata ruumist, kus nad asuvad. Pärast andmete Markovi graafiks teisendamist on võimalik iga kahe graafi tipu vahel defineerida difusioonkaugus, kasutades juhuslikku ekslemist nende tippude vahel. Markovi üleminekumaatriksi omaväärtuste ja omavektorite abil saame defineerida difusioonikaardid, mis kujutavad graafi eukleidilisse ruumi. Selle töö peaeesmärk on tõestada teoreem, mis väidab, et difusioonkaugus graafil on konstandiga korrutamise täpsuseni võrdne eukleidilise kaugusega kujutusruumis. Me illustreerime seda teoreemi kasutades nii sünteetilisi kui ka juba tuntud andmehulki.

CERCS teaduseriala: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria).

Märksõnad: difusioonikaardid, difusioonkaugus, spektraalteoreem, Markovi ahelad, mõõtmete vähendamine.

Introduction

In statistical data analysis and computer vision, we are often interested in focusing only on the intrinsic geometry of the data, not on its location in the ambient space it lives in. In data analysis, understanding the intrinsic geometry can be used to capture nonlinear patterns in the data, and, as a result, it is sometimes possible to represent the most prominent features of the dataset in a lower-dimensional space. In computer vision, understanding the intrinsic geometry can be used for solving correspondence problems.

Isolating the data's position from its local connectivity is a non-trivial task. One way to do that is to turn the dataset of interest into a weighted graph using kernels that satisfy certain conditions. After normalising the weighted graph, we can embed it using functions called diffusion maps. Diffusion maps capture local connectivity in the data by analysing how random walks move between vertices in the graph.

The variant of diffusion maps that we study in this thesis was presented by Nadler, Lafon, Coifman and Kevrekidis in 2005 (see [Nad+05]). They proved that, after turning the dataset into a weighted graph via the Gaussian kernel and embedding it via diffusion maps, the diffusion distance between any two vertices on the graph is proportional to the Euclidean distance in the embedded space. In the paper, the authors do not analyse in detail what conditions the kernel must satisfy for the embedding to work as expected. They also prove the main theorem, but rely on auxiliary results which are often not stated. Our aim is to prove the theorem in a self-contained manner and to identify sufficient conditions on the kernel for the theorem to be applicable.

The thesis is organised as follows. Section 1 explains the basic terms used in the thesis and contains key results used in later proofs. Section 2 defines diffusion distance and diffusion maps, and proves that the diffusion distance is proportional to the Euclidean distance in the embedded space. We begin by providing a complete proof of the spectral theorem for real symmetric matrices, since it is a key theorem we use later to define diffusion maps. After that, we build on [Nad+05], but we prove the theorem in a self-contained way. We follow the finite Markov-matrix formulation of [Nad+05] and make explicit a set of assumptions on the kernel function under which the proof goes through. We also modify the proof since some of the steps in [Nad+05] were slightly ambiguous.

Section 3 applies the theory to both synthetic and established datasets. We illustrate the key properties of diffusion maps and show how they can be used for solving correspondence problems. We also illustrate how using diffusion maps differs from classical linear dimensionality reduction techniques. We use principal component analysis (see [Jol02, Section 1.1]) as an example, since it is a well-known linear method. We finish by applying diffusion maps to the correspondence problem on figures from the Dynamic FAUST dataset (see [Bog+17]). Since the dataset contains non-rigid transformations, the embeddings do not align as cleanly as in the synthetic examples. We nonetheless believe the example is valuable, both as a demonstration using real data and as motivation for the discussion of what conditions must be satisfied for diffusion maps to work to their full potential. The code used to generate plots and calculate the embeddings in Section 3 can

be found on GitHub (see [Nur26]).

Large language models by Google (see [Goo26]), OpenAI (see [Ope25]), and Anthropic (see [Ant26]) were used during the writing of this thesis to generate alternative ideas, better understand materials used, find relevant literature and references, find mistakes in proofs, help write and debug the Python code for creating plots and calculations, and improve the readability of existing text. None of the content in the thesis was written by artificial intelligence, and its output was thoroughly checked.

1 Definitions and Preliminaries

In this section, we will introduce essential definitions and results required in the later sections of this thesis. Table 1 summarises the general notation used throughout the thesis.

Table 1: General notation used throughout the thesis.

Notation	**Meaning**
$\mathrm{Mat}_n(\mathbb{K})$	Square $n \times n$ matrix over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$
$\mathrm{Mat}_{n\times 1}(\mathbb{K})$	Column vectors of length n with elements in $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$
A_{ij}	(ij)-th element of a matrix $A \in \mathrm{Mat}_n(\mathbb{K})$
$[v_1, \ldots, v_n]$	Matrix with columns $v_1, \ldots, v_n \in \mathrm{Mat}_{n\times 1}(\mathbb{K})$
e_k	k-th canonical basis vector
Θ_n	Zero matrix of shape $n \times n$
Θ	Zero matrix (with dimensions clear from context)
I_n	Identity matrix of shape $n \times n$
$\mathbf{1}$	Vector $(1, \ldots, 1)^T$
$\|\cdot\|_2$	Euclidean 2-norm
$\mathrm{card}(X)$	Cardinality of set X
δ_{ij}	Kronecker delta

1.1 Algebra

We start with some general definitions from linear algebra; later in this subsection, we will also prove some key results used in the following sections of the thesis.

Definition 1.1. Let $A \in \mathrm{Mat}_n(\mathbb{K})$ be a square matrix. Vectors $\psi \in \mathrm{Mat}_{n\times 1}(\mathbb{C}) \setminus \{0\}$ and $\varphi \in \mathrm{Mat}_{n\times 1}(\mathbb{C}) \setminus \{0\}$ are called respectively *right* and *left eigenvectors* if

$$A\psi = \lambda\psi \text{ and } \varphi^T A = \mu\varphi^T,$$

where $\lambda, \mu \in \mathbb{C}$ are eigenvalues.

Note that in Definition 1.1, eigenvalues of real matrices can be either real or complex.

Definition 1.2. Let $M \in \mathrm{Mat}_n(\mathbb{R})$. The value

$$\rho(M) \coloneqq \max\{|\lambda| : \lambda \text{ is an eigenvalue of } M\}$$

is called the *spectral radius of* M.

Since almost all the matrices we work with in this thesis are square, we will define some useful types of square matrices.

Definition 1.3. Let $A \in \mathrm{Mat}_n(\mathbb{R})$.

1. Matrix A is called *diagonal* if $A_{ij} = 0$ for every $i \neq j$ where $i, j \in \{1, \ldots, n\}$.
2. Matrix A is called *orthogonal* if its transpose is equal to its inverse: $A^T = A^{-1}$.
3. Matrix A is called *symmetric* if it is equal to its transpose: $A = A^T$.
4. A symmetric matrix A is called *positive semi-definite* (PSD, for short) if for all vectors $x \in \mathbb{R}^n \setminus \{0\}$ one has $x^T A x \geqslant 0$.

The following theorem gives us a useful property of positive semi-definite matrices.

Theorem 1.4 (see [HJ13, Theorem 7.2.1]). *Let $A \in Mat_n(\mathbb{R})$ be symmetric. The following statements are equivalent:*

(i) A is positive semi-definite;

(ii) all eigenvalues of A are non-negative.

Diagonal matrices have many nice properties, especially if all their diagonal entries are positive. Given a diagonal matrix $D \in \mathrm{Mat}_n(\mathbb{R})$ with positive entries on its diagonal, meaning $D_{ii} > 0$, for $i \in \{1, \ldots, n\}$, we can denote by D^t ($t \in \mathbb{R}$) a diagonal matrix with $\left(D^t\right)_{ii} = (D_{ii})^t$.

An important result in linear algebra is called the Gram-Schmidt orthogonalisation process, but before we can state it, we must first define inner product spaces.

Definition 1.5. Let H be a linear space over the field $\mathbb{R}$ and $(a, b) \in H^2$ be an ordered pair. The mapping $H^2 \ni (a, b) \mapsto \langle a, b\rangle \in \mathbb{R}$ is called an *inner product* if the following conditions are met for all $a, b, c \in H$ and $k \in \mathbb{R}$:

1. $\langle a, a\rangle \geqslant 0$ and $[\langle a, a\rangle = 0 \iff a = 0]$;
2. $\langle a, b\rangle = \langle b, a\rangle$;
3. $\langle a + b, c\rangle = \langle a, c\rangle + \langle b, c\rangle$;
4. $\langle ka, b\rangle = k\langle a, b\rangle$.

A space E equipped with an inner product $\langle \cdot, \cdot \rangle$ is called an *inner product space*.

Definition 1.6. Let E be an inner product space. Vectors $v_1, v_2 \in E$ are called *orthogonal* if $\langle v_1, v_2\rangle = 0$. A system of vectors $\{v_1, \ldots, v_n\} \subset E$ is called orthogonal if they are pair-wise orthogonal: for all $i, j \in \{1, \ldots, n\}$, if $i \neq j$ then $\langle v_i, v_j\rangle = 0$.

If additionally for every $v_i \in \{v_1, \ldots, v_n\}$ the length $\|v_i\| \coloneqq \sqrt{\langle v_i, v_i\rangle} = 1$ then the system is called *orthonormal*.

Now we can state the Gram-Schmidt orthogonalisation process.

Theorem 1.7 (Gram-Schmidt orthogonalisation process; see [Str16, Section 4.4]). *Given a finite or countable linearly independent system of vectors in an inner product space H, the Gram-Schmidt process produces an orthogonal system of non-zero vectors that spans the same subspace. After normalising these vectors, we obtain an orthonormal system which spans the same subspace.*

Proposition 1.8 (Existence of eigenvalues; see [Axl24, Theorem 5.19]). *Every operator*[1] *on a finite-dimensional nonzero complex vector space has an eigenvalue.*

Since matrices are linear operators on finite-dimensional vector spaces, it follows directly from Proposition 1.8 that every matrix $A \in \mathrm{Mat}_n(\mathbb{R})$ has at least one eigenvalue $\lambda \in \mathbb{C}$.

Proposition 1.9 (see [Laa23, Proposition 11.23]). *Let $A \in Mat_n(\mathbb{R})$. Then the following statements are equivalent:*

(i) A is orthogonal;

(ii) the columns of A form an orthonormal system.

We will end this subsection by proving that all symmetric real matrices have real eigenvalues. For a complex number $\lambda \in \mathbb{C}$, we denote its complex conjugate by $\overline{\lambda}$.

Lemma 1.10. *If $A \in Mat_n(\mathbb{R})$ is symmetric, then its eigenvalues are real.*

Proof. Let $A \in \mathrm{Mat}_n(\mathbb{R})$ be a symmetric matrix. By Proposition 1.8, the matrix A has at least one eigenvalue. Suppose $\lambda \in \mathbb{C}$ is an eigenvalue. We will show that $\lambda = \overline{\lambda}$. First notice that since A is real, the following holds:

$$Av = \lambda v \Leftrightarrow \overline{Av} = \overline{\lambda v} \Leftrightarrow A\overline{v} = \overline{\lambda}\overline{v}.$$

Now, denote the corresponding eigenvector of eigenvalue λ as v. This means $Av = \lambda v$. On one hand

$$v^T A\overline{v} = (Av)^T\,\overline{v} = (\lambda v)^T\overline{v} = \lambda v^T\overline{v}.$$

On the other hand

$$v^T A\overline{v} = v^T\overline{\lambda}\overline{v} = \overline{\lambda}v^T\overline{v}.$$

Hence $\lambda v^T\overline{v} = \overline{\lambda}v^T\overline{v} \implies \lambda = \overline{\lambda}$. Note that we can cancel out $v^T\overline{v} = |v_1|^2 + \cdots + |v_n|^2 \neq 0$ because v is an eigenvector. Therefore, we have proven that all eigenvalues of A are real. ∎

The last theorem we state in this subsection is called the Perron-Frobenius theorem.

[1] Here, the operator is a linear map from a vector space to itself.

Theorem 1.11 (Perron-Frobenius Theorem, see [HJ13, Theorem 8.2.2]). *Let $A \in Mat_n(\mathbb{R})$ be a positive matrix. Then the eigenvector corresponding to the largest eigenvalue can be chosen to be positive, which means there exists an eigenvector v such that*

$$Av = \rho(A)v,$$

and $v_i > 0$ for every $i \in \{1, \dots, n\}$.

1.2 Probability theory

To formally define diffusion distance and diffusion maps, we first introduce Markov chains and state some theorems about random processes. We start with some general definitions from probability. This section is strongly inspired by [Axl20].

Definition 1.12. Let $\mathcal{F}$ be a σ-algebra defined on a set Ω and let $\boldsymbol{P}$ be a measure on $(\Omega, \mathcal{F})$. The measure $\boldsymbol{P}$ is called a *probability measure* if $\boldsymbol{P}(\Omega) = 1$. In this case, the triple $(\Omega, \mathcal{F}, \boldsymbol{P})$ is called a *probability space*. Set Ω is called a *sample space*, elements of the σ-algebra $\mathcal{F}$ are called *events*, and for $A \in \mathcal{F}$ the value $\boldsymbol{P}(A)$ is called the *probability of A*.

From now on, let $(\Omega, \mathcal{F}, \boldsymbol{P})$ be a probability space.

As Definition 1.12 suggests, the sample space can be quite abstract, for example, the set of all possible trajectories of a stochastic process. To work with such structures, it is often useful to map the events to real numbers using functions called random variables.

Definition 1.13. Let E be a finite set. A function $X : \Omega \to E$ such that $\{\omega \in \Omega : X(\omega) = s\} \in \mathcal{F}$ for every $s \in E$ is called a *random variable*.

Let $X : \Omega \to E$ be a random variable and let $s \in E$. For the probability that X takes the value s, we write

$$\boldsymbol{P}(X = s) \coloneqq \boldsymbol{P}(\{\omega \in \Omega : X(\omega) = s\}).$$

More generally, for a finite collection of random variables $X_0, \dots, X_{n-1}$ and $s_0, \dots, s_{n-1} \in E$ we write

$$\boldsymbol{P}(X_0 = s_0, \dots, X_{n-1} = s_{n-1}) \coloneqq \boldsymbol{P}(\{\omega \in \Omega : X_0(\omega) = s_0, \dots, X_{n-1}(\omega) = s_{n-1}\}).$$

Definition 1.14. Let $A, B \in \mathcal{F}$ be events with $\boldsymbol{P}(B) > 0$. The *conditional probability of A given B* is

$$\boldsymbol{P}(A \mid B) \coloneqq \frac{\boldsymbol{P}(A \cap B)}{\boldsymbol{P}(B)}.$$

Now we can introduce discrete-time stochastic processes on a finite state space.

Definition 1.15. Let E be a finite set. A sequence $\{X_t\}_{t \geqslant 0}$, where each $X_t : \Omega \to E$ is a random

variable, is called a *discrete-time stochastic process*, and E is called its *state space*. Elements of E are called *states*, and we say the process is in *state s at time t* if $X_t = s$.

A special type of discrete-time stochastic process is called a Markov chain.

Definition 1.16. Let $\{X_t\}_{t\geqslant 0}$ be a discrete-time stochastic process with state space E. If for all $n \in \mathbb{N} \cup \{0\}$ and all states $s_0, \ldots, s, l \in E$ the following property

$$\boldsymbol{P}(X_{n+1} = l \mid X_n = s, \ldots, X_0 = s_0) = \boldsymbol{P}(X_{n+1} = l \mid X_n = s)$$

is satisfied, then the discrete-time stochastic process is called a *Markov chain*. This property is called the *Markov property*. The chain is called *time-homogeneous* if, additionally, the right-hand side of the property does not depend on n.

Definition 1.17. Let $\{X_t\}_{t\geqslant 0}$ be a time-homogeneous Markov chain, let $E = \{s_0, \ldots, s_{n-1}\}$ be its state space. The matrix $M \in \mathrm{Mat}_n(\mathbb{R})$, with $M_{ij} \coloneqq \boldsymbol{P}(X_1 = s_j \mid X_0 = s_i)$, is called the *transition matrix* of the Markov chain. Since each row of matrix M is a probability distribution over E, all its entries are non-negative, and each row sums up to 1. Square matrices with these two properties satisfied are called *stochastic matrices*.

Example 1.18. Figure 1 contains an example of a time-homogeneous Markov chain with state space $E = \{s_0, s_1\}$ and with transition probability α from state s_0 to s_1, and β from state s_1 to s_0.

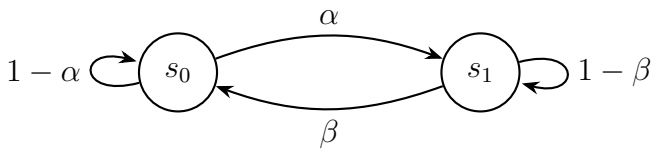


Figure 1: A time-homogeneous Markov chain with two states, and transition probabilities $\alpha, \beta \in [0, 1]$.

The corresponding Markov transition matrix is

$$M = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

A useful property of time-homogeneous Markov chain transition matrices is called the *Chapman-Kolmogorov equation*. First, a quick remark regarding notation. For $t \in \mathbb{N} \cup \{0\}$ we denote the probability of moving from state s to state l in t steps by $p(t, l \mid s) \coloneqq \boldsymbol{P}(X_t = l \mid X_0 = s)$. By the Chapman-Kolmogorov equation (see [GS20, Theorem (7), §6.1]), the entry with row corresponding to state s and column corresponding to state l of the Markov transition matrix M^t equals $p(t, l \mid s)$.

States of a Markov chain can have different properties. The next definition provides insight into some state properties we will use in the proofs that follow.

Definition 1.19. Let $s, l \in E$ be states of a Markov chain. State l is said to be *accessible* from

state s if $p(t, l \mid s) > 0$ for some $t \in \mathbb{N} \cup \{0\}$, in this case we write $s \to l$. States s and l are said to *communicate* if $s \to l$ and $l \to s$, we denote this by $s \leftrightarrow l$. A Markov chain is called *irreducible* if every pair of states communicates.

Another useful property of a state is its periodicity.

Definition 1.20. Let $s \in E$ be a state. The greatest common divisor (gcd, for short) of the set $\{t \in \mathbb{N} : p(t, s \mid s) > 0\}$ (with the period defined as ∞ if the set is empty) is called the *period* of the state. The Markov chain is called *aperiodic* if every state has period 1.

Often we are interested in how the Markov chain performs over many steps. It turns out that irreducibility and aperiodicity together guarantee that the process behaves "nicely". We will now state a key theorem regarding that.

Theorem 1.21 (The Fundamental Theorem of Markov chains; see [VW25, Theorem (1)]). *Let $\{X_t\}_{t \geqslant 0}$ be a Markov chain over a finite state space, with transition matrix M. Suppose the chain is irreducible and aperiodic. Then the following hold:*

1. *There exists a unique stationary distribution $\pi = (\pi_0, \ldots, \pi_{n-1})$ over the states such that for any states s, l*
$$\lim_{t \to \infty} \boldsymbol{P}(X_t = s \mid X_0 = l) = \pi_s.$$
2. *π is a left eigenvector of matrix M, with eigenvalue 1, meaning $\pi M = \pi$.*

Theorem 1.22 (see [HJ13, Theorem 8.4.4]). *Let $M \in Mat_n(\mathbb{R})$, $n \geqslant 2$, be irreducible and non-negative. Then $\rho(M)$ is an algebraically simple eigenvalue of M, that is, it has algebraic multiplicity 1.*

Note that the result from Theorem 1.22 holds for any 1×1 real matrix $m \in \mathrm{Mat}_1(\mathbb{R})$ too. Indeed, by solving for eigenvalues $\det(m - \lambda I) = 0 \Leftrightarrow m = \lambda$ and thus λ has algebraic multiplicity 1.

2 Diffusion Distance and Diffusion Maps

Many datasets can be modelled as graphs, often as fully connected, weighted graphs where edge weights reflect the similarity between data points. The diffusion distance provides a natural way to measure the similarity between vertices in such graphs, but computing it directly can be computationally expensive. In this section, we prove that under sufficient conditions on the kernel, diffusion maps produce an embedding in which Euclidean distance is equal, up to a scaling constant, to diffusion distance in the original space. While this equivalence, by itself, does not reduce computational cost, in practice, it is often sufficient to use diffusion maps to embed the data into a lower-dimensional Euclidean space, in which case the complexity decreases significantly. We will not provide rigorous mathematical justification for using diffusion maps for dimensionality reduction, but we will illustrate their efficiency in Section 3.

Before we formally define diffusion distance and prove how to embed the graph into Euclidean space, we need to prove the spectral theorem.

2.1 Spectral Theorem

Since in this subsection we will frequently work with matrices that have a block structure, we start by defining the block decomposition of matrices. This will make the proofs in this subsection easier to follow.

Definition 2.1. Let $A \in \mathrm{Mat}_{m\times n}(\mathbb{R})$. We say that the matrix A has a *block decomposition* if

$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right)$$

where $A_{ij} \in \mathrm{Mat}_{m_i\times n_j}(\mathbb{R})$ such that $m_1 + m_2 = m$, $n_1 + n_2 = n$ and $i, j \in \{1, 2\}$.

We will now state some useful properties of block decomposed matrices.

Property 2.2. *Let $A \in Mat_{m\times n}(\mathbb{R})$ and let*

$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right)$$

be its block decomposition. Then

$$A^T = \left(\begin{array}{c|c} A_{11}^T & A_{21}^T \\ \hline A_{12}^T & A_{22}^T \end{array}\right).$$

Property 2.3 (see [Str16, Block multiplication]). *Let A and B be matrices with block decompositions*

$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right),\ B = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right).$$

If blocks of A can multiply blocks of B, then

$$AB = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right)\left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right) = \left(\begin{array}{c|c} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ \hline A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{array}\right).$$

The main result we want to prove in this subsection is the spectral theorem, which states that every n-dimensional symmetric matrix has real eigenvalues and its eigenvectors form an orthonormal basis of $\mathbb{R}^n$. For better readability, we will provide the proof split into smaller lemmas.

Lemma 2.4. *Let $A \in Mat_n(\mathbb{R})$ be a symmetric matrix. Then there exists an orthogonal matrix $R \in Mat_n(\mathbb{R})$ such that $R^{-1}AR$ is diagonal.*

Proof. We will prove it by induction.

The base case: if $A \in \mathrm{Mat}_1(\mathbb{R})$ the statement holds, because we can choose $\mathrm{Mat}_1(\mathbb{R}) \ni R = (1)$.

The inductive step: we now assume the statement holds for $(n-1)$-dimensional matrices.

We will now show that the statement holds for n-dimensional matrices: let $A \in \mathrm{Mat}_n(\mathbb{R})$ be symmetric. Since every finite-dimensional matrix has at least one eigenvalue (see Proposition 1.8), we can denote the eigenvalue of A as $\lambda_1 \in \mathbb{C}$. Since A is real and symmetric, from Lemma 1.10, we know that $\lambda_1 \in \mathbb{R}$. Note that since A is a real matrix and the eigenvalue λ_1 is real, then the corresponding eigenvector v_1 is also real. Since eigenvectors are defined up to normalisation, we can choose v_1 such that $\|v_1\|_2 = 1$. Since v_1 is a non-zero vector in $\mathbb{R}^n$, it can be extended to a basis $\{v_1, u_2, \ldots, u_n\}$ of $\mathbb{R}^n$. Applying the Gram-Schmidt orthogonalisation process (see Theorem 1.7) to this basis and normalising the resulting vectors, we end up with an orthonormal basis $\{v_1, v_2, \ldots, v_n\}$ of $\mathbb{R}^n$, where the first vector remains v_1 since it was already a unit vector. Now let us define matrix $P = [v_1, v_2, \ldots, v_n]$. Because the basis vectors are orthonormal, we know by Proposition 1.9 that P is orthogonal. Let $B = P^{-1}AP$. We will show that B is symmetric:

B is symmetric: since the matrix P is orthogonal we get $B = P^TAP$ therefore

$$B^T = \left(P^T(AP)\right)^T = P^TA^TP = P^TAP.$$

Consider Be_1, where e_1 is the first standard basis vector of $\mathbb{R}^n$. We compute $P^TAPe_1 = Be_1$, which equals the first column of the matrix B. On the other hand,

$$P^TAPe_1 = P^TAv_1 = P^T\lambda_1 v_1 = \lambda_1 P^T v_1 \overset{(*)}{=} \lambda_1 e_1.$$

Equality $(*)$ holds, because P^T is a matrix with vectors $v_1, v_2, \ldots, v_n$ as rows, since they are orthonormal $v_i^T v_j = \delta_{ij}$. So we get

$$B = \left(\begin{array}{c|c} \lambda_1 & \Theta \\ \hline \Theta & C \end{array}\right),$$

where matrix $C \in \mathrm{Mat}_{n-1}(\mathbb{R})$. Note that C is also symmetric. Indeed, this follows directly from B being a symmetric matrix. By our inductive hypothesis, there exists a matrix Q such that the matrix $D := Q^{-1}CQ$ is diagonal and the matrix Q is orthogonal. Now, let us define the matrix R as follows:

$$R := P \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q \end{array}\right).$$

Observe that

(1) R is orthogonal:

$$R^{-1} = \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q^{-1} \end{array}\right) P^{-1} = \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q^{T} \end{array}\right) P^{T} = R^{T}.$$

Note that the inverse of R is found correctly, since

$$\left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q^{-1} \end{array}\right)\left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q \end{array}\right) = \left(\begin{array}{c|c} 1\cdot 1 + \Theta\cdot\Theta & 1\cdot\Theta + \Theta Q \\ \hline \Theta\cdot 1 + Q^{-1}\cdot\Theta & \Theta\cdot\Theta + Q^{-1}Q \end{array}\right) = \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & I \end{array}\right) = I.$$

(2) $R^{-1}AR$ is diagonal:

$$\begin{aligned} R^{-1}AR = R^{T}AR &= \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q^{T} \end{array}\right) P^{T}AP \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q \end{array}\right) \\ &= \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q^{T} \end{array}\right) B \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q \end{array}\right) \\ &= \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q^{T} \end{array}\right)\left(\begin{array}{c|c} \lambda_1 & \Theta \\ \hline \Theta & C \end{array}\right)\left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & Q \end{array}\right) \\ &= \left(\begin{array}{c|c} \lambda_1 & \Theta \\ \hline \Theta & Q^{T}CQ \end{array}\right) = \left(\begin{array}{c|c} \lambda_1 & \Theta \\ \hline \Theta & D \end{array}\right) \end{aligned}$$

and matrix D is diagonal by assumption.

Hence, R is orthogonal and $R^{-1}AR$ is diagonal, which completes the inductive step. ■

Proposition 2.5. *Let $A \in Mat_n(\mathbb{R})$. The following statements are equivalent:*

(i) There exists an orthogonal matrix R such that the matrix $R^{-1}AR$ is diagonal.

(ii) $\mathbb{R}^n$ has an orthonormal basis of eigenvectors of A.

Proof. Firstly, assume there exists an orthogonal matrix $R \in \mathrm{Mat}_n(\mathbb{R})$ such that $R^{-1}AR$ is

diagonal with elements $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ on the diagonal. Multiplying both sides by R we get

$$AR = R \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Let $v_1, \ldots, v_n$ denote the columns of R. Therefore, the j-th column of matrix AR is Av_j and the j-th column of the right-hand side is $\lambda_j v_j$. Comparing the columns we get $Av_j = \lambda_j v_j$ for every $j \in \{1, \ldots, n\}$ and thus each v_j is an eigenvector of A with eigenvalue λ_j. From the orthogonality of R, its columns form an orthonormal system in $\mathbb{R}^n$ (see Proposition 1.9). Since there are n of them, they form an orthonormal basis in $\mathbb{R}^n$, which is exactly what we needed to show.

Assume now that eigenvectors $\{v_1, \ldots, v_n\}$ of a matrix A form an orthonormal basis of $\mathbb{R}^n$. Denote the corresponding eigenvalues with $\{\lambda_1, \ldots, \lambda_n\}$. Let R be a matrix with eigenvectors $v_1, \ldots, v_n$ for its columns. Since the eigenvectors form an orthonormal basis, R is orthogonal (see Proposition 1.9). Now, computing the matrix AR column-by-column, we get

$$AR = [Av_1, \ldots, Av_n] = [\lambda_1 v_1, \ldots, \lambda_n v_n] = R \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Multiplying both sides by R^{-1} from the left we get

$$R^{-1}AR = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

which is diagonal. Therefore, we have shown that the statements are indeed equivalent. ■

Theorem 2.6 (Spectral Theorem, see [Str16, Spectral Theorem, p. 339]). *Let $A \in Mat_n(\mathbb{R})$ be symmetric. Then the following two statements hold:*

1. *A has real eigenvalues.*
2. *There exists an orthonormal basis of $\mathbb{R}^n$ which consists of eigenvectors of A.*

Proof.

1. Follows directly from Lemma 1.10.
2. By Lemma 2.4, there exists an orthogonal matrix $R \in \mathrm{Mat}_n(\mathbb{R})$ such that $R^{-1}AR$ is diagonal. By Proposition 2.5, the existence of such a matrix R is equivalent to eigenvectors of A forming an orthonormal basis in $\mathbb{R}^n$, which is exactly what we needed to show.

■

2.2 Diffusion Theory

In this subsection, we will define diffusion distance and diffusion maps in a self-contained manner, building on [Nad+05]. The main result of this thesis, Theorem 2.13, states that the diffusion distance on the graph is equal, up to a known scaling constant, to the Euclidean distance in the embedded space. Our presentation differs from [Nad+05] mainly in two ways. Firstly, we provide a self-contained proof, while [Nad+05] relies on many auxiliary results. Secondly, we state sufficient conditions on the kernel under which the theorem is applicable; this is not done in [Nad+05].

Although we have stated and proved most of the necessary results, we still need to prove some additional ones. Therefore, to make this subsection easier to follow, we begin by providing a roadmap outlining the key checkpoints. After that, we will start from the beginning.

(I). We start with a finite one- or higher-dimensional set of data, which we denote by $\mathcal{X}$. We can think of the points of $\mathcal{X}$ as vertices of an undirected weighted graph $G = (\mathcal{X}, W)$ with W being the set of weighted edges, which we will define shortly, such that we can construct a Markov chain on the graph. Note that the elements of $\mathcal{X}$ have an arbitrary but fixed order: the labelling $x_0, x_1, \ldots, x_{n-1}$ is chosen arbitrarily once and kept consistent throughout the subsection. This also justifies the notation $v_i(x_j)$ for denoting the coordinate of vector v_i corresponding to the position of data point x_j.

(II). We will define diffusion distance at time $t \in \mathbb{N} \cup \{0\}$ as

$$\begin{aligned} D_t^2(x_i, x_j) &\coloneqq \|p(t, x|x_i) - p(t, x|x_j)\|_w^2 \\ &\coloneqq \sum_{x \in \mathcal{X}} \left(p(t, x|x_i) - p(t, x|x_j)\right)^2 w(x), \end{aligned}$$

where $x_i, x_j \in \mathcal{X}$, $p(t, x \mid x_i)$ denotes the probability of a random walk landing at location x at time t, given a starting location x_i at time $t = 0$, and $w(x)$ is a weight. We then define the diffusion maps as

$$\Psi_t : \mathcal{X} \ni x_i \mapsto \left(\lambda_1^t \psi_1(x_i), \ldots, \lambda_k^t \psi_k(x_i)\right) \in \mathbb{R}^k,$$

where $\mathbb{N} \ni k \leqslant \operatorname{card}(\mathcal{X}) - 1$, λ_i is the eigenvalue, and ψ_i is the corresponding right eigenvector of the Markov matrix defined on the graph G.

(III). We will finish the subsection by proving that diffusion distance on the graph is equal, up to a scaling factor, to the Euclidean distance in the embedded space, where the embedding uses all $(\operatorname{card}(\mathcal{X}) - 1)$ non-trivial eigenvectors:

$$D_t^2(x_i, x_l) = \frac{1}{c} \sum_{j=1}^{\operatorname{card}(\mathcal{X})-1} \left(\lambda_j^t\right)^2 \left(\psi_j(x_i) - \psi_j(x_l)\right)^2 = \frac{1}{c} \left\|\Psi_t(x_i) - \Psi_t(x_l)\right\|^2,$$

where $x_i, x_l \in \mathcal{X}$ and $c \in \mathbb{R}$.

We now begin by denoting the set of data $\mathcal{X} \coloneqq \{x_0, \ldots, x_{n-1}\} \subset \mathbb{R}^p$ $(n, p \in \mathbb{N})$. To build the

undirected weighted graph $G = (\mathcal{X}, W)$, with $W \subset \mathcal{X} \times \mathcal{X}$, we need to determine the weights. For that, we start by defining a kernel function.

Definition 2.7. A function $k : \mathcal{X} \times \mathcal{X} \to (0, \infty)$ is called a *kernel function* (kernel, for short).

In general, kernels need not have a strictly positive range, but in our case, it is important for the theory to hold.

Additionally, we assume that the kernel k is

1. Positive semi-definite:
$$\sum_{i,j=0}^{m-1} c_i c_j k(x_i, x_j) \geqslant 0$$
for any $x_0, \ldots, x_{m-1} \in \mathcal{X}$, $c_0, \ldots, c_{m-1} \in \mathbb{R}$ ($m \in \mathbb{N}$).
2. Symmetric: $k(x_i, x_j) = k(x_j, x_i)$ for every $x_i, x_j \in \mathcal{X}$.

Example 2.8. Let $x_i, x_j \in \mathbb{R}^p$ and $\varepsilon > 0$. The *Gaussian kernel* is a strictly positive, symmetric, positive semi-definite kernel
$$k(x_i, x_j) \coloneqq \exp\left(-\frac{\|x_i - x_j\|^2}{2\varepsilon}\right).$$

Here $\varepsilon > 0$ can be thought of as the width or radius, since for a large ε the kernel will evaluate to near 1 even for points that are far apart.

Now we can determine the weights for the graph G by defining $W \in \mathrm{Mat}_n(\mathbb{R})$ as $W_{ij} \coloneqq k(x_i, x_j)$ ($i, j \in \{0, \ldots, n-1\}$). We define the degree of a vertex as the sum of weights connecting it to other vertices. This allows us to naturally define the random-walk normalisation matrix D.

Definition 2.9. Let $W, D \in \mathrm{Mat}_n(\mathbb{R})$. The diagonal matrix D is called the *random-walk normalisation matrix* if
$$D_{ii} \coloneqq \sum_{j=0}^{n-1} W_{ij}.$$

Since, in our case, there are no isolated vertices (k is a strictly positive kernel), the matrices D^{-1} and $D^{-1/2}$ are well-defined.

We now have all the necessary components to define a Markov chain on the graph G. We construct a time-homogeneous Markov chain $\{X_t\}_{t \geqslant 0}$ with the vertices $\mathcal{X}$ as its states and the transition probability between the states specified via the normalised weighted edges:
$$\forall t \in \mathbb{N} \cup \{0\} : \ \boldsymbol{P}(X_{t+1} = x_j \mid X_t = x_i) = \frac{k(x_i, x_j)}{\sum_{m=0}^{n-1} k(x_i, x_m)}.$$

Arranging these probabilities into an n-dimensional matrix $M \in \mathrm{Mat}_n(\mathbb{R})$, such that

$$M_{ij} = \boldsymbol{P}(X_{t+1} = x_j \mid X_t = x_i),$$

we have constructed a Markov transition matrix.

It turns out that $M = D^{-1}W$. Indeed,

$$\left(D^{-1}W\right)_{ij} = \sum_{k=0}^{n-1} D^{-1}_{ik} W_{kj} = \frac{W_{ij}}{D_{ii}} = \frac{W_{ij}}{\sum\limits_{m=0}^{n-1} W_{im}} = \frac{k(x_i, x_j)}{\sum\limits_{m=0}^{n-1} k(x_i, x_m)} = M_{ij}.$$

Example 2.10. Let W be a weight matrix such that

$$W = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}.$$

Then the corresponding random walk normalisation matrix is

$$D = \begin{pmatrix} 1.25 & 0 \\ 0 & 1.25 \end{pmatrix},$$

and the resulting row-stochastic Markov transition matrix $M = D^{-1}W$ is

$$M = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}.$$

We will now define a matrix $D^{1/2}MD^{-1/2} =: M_s \in \mathrm{Mat}_n(\mathbb{R})$. Note that M_s is symmetric:

$$M_s = D^{1/2}MD^{-1/2} = D^{1/2}D^{-1}WD^{-1/2} = D^{-1/2}WD^{-1/2},$$

and on the other hand

$$M_s^T = \left(\left(D^{-1/2}W\right)D^{-1/2}\right)^T = \left(D^{-1/2}\right)^T W^T \left(D^{-1/2}\right)^T = D^{-1/2}WD^{-1/2}$$

because D and W are symmetric by construction. Additionally, M is similar to M_s:

$$M_s = D^{1/2}M\left(D^{1/2}\right)^{-1}.$$

This means that M and M_s share eigenvalues. Since M_s is symmetric, from the Spectral Theorem (see Theorem 2.6) we know that the eigenvalues $\{\lambda_j\}_{j=0}^{n-1}$ are all real. Denote the corresponding eigenvectors of M_s by $\{v_j\}_{j=0}^{n-1}$. Note that they form an orthonormal basis in $\mathbb{R}^n$, which also means that $\|v_j\|_2 = 1$. Additionally, we choose v_0 to be positive; this is possible by Theorem 1.11 because M_s is a positive matrix.

Additionally, note that M_s is PSD. Indeed, for $x \in \mathbb{R}^n \setminus \{0\}$ denote $y = D^{-1/2}x$ and therefore $y^T = x^T D^{-1/2}$. Now

$$x^T M_s x = x^T D^{-1/2} W D^{-1/2} x = y^T W y \geqslant 0$$

because W is, by assumption, PSD.

Now, it turns out that the left and right eigenvectors of M are respectively

$$\varphi_j = D^{1/2} v_j \text{ and } \psi_j = D^{-1/2} v_j.$$

Indeed:

$$\begin{aligned}\varphi_j^T M &= v_j^T D^{1/2} M \left(D^{-1/2} D^{1/2}\right) = v_j^T M_s D^{1/2} = (M_s v_j)^T D^{1/2} \\ &= (\lambda_j v_j)^T D^{1/2} = \lambda_j \left(D^{1/2} v_j\right)^T = \lambda_j \varphi_j^T\end{aligned}$$

and

$$M\psi_j = MD^{-1/2} v_j = D^{-1/2} D^{1/2} M D^{-1/2} v_j = D^{-1/2} \lambda_j v_j = \lambda_j \psi_j.$$

Since the vectors v_j are orthonormal in $\mathbb{R}^n$, the following holds:

$$\langle \psi_i, \varphi_j \rangle = v_i^T D^{-1/2} D^{1/2} v_j = \delta_{ij}.$$

Therefore $\{\psi_j\}_{j=0}^{n-1}$ and $\{\varphi_j\}_{j=0}^{n-1}$ are biorthonormal.

Before continuing, we will prove that strictly positive Markov matrices have unique stationary distributions.

Theorem 2.11. *Let $M \in Mat_n(\mathbb{R})$ be a Markov transition matrix with strictly positive entries, meaning $M_{ij} > 0$ for every $i, j \in \{0, \ldots, n-1\}$, then M has a unique stationary distribution.*

Proof. Let $M \in \mathrm{Mat}_n(\mathbb{R})$ be a strictly positive Markov transition matrix. We will first show that M is aperiodic and then that M is irreducible.

1. We want to show that for any $i \in \{0, \ldots, n-1\}$ the greatest common divisor of the set $\{t \in \mathbb{N} : p(t, i \mid i) > 0\}$ is 1. Since M is strictly positive $p(1, i \mid i) > 0$, therefore $1 \in \{t \in \mathbb{N} : p(t, i \mid i) > 0\}$ which is what we wanted to show.
2. We want to show that $p(t, j \mid i) > 0$ for every $i, j \in \{0, \ldots, n-1\}$ and for some $t \in \mathbb{N}$. Since the Markov transition matrix is strictly positive $p(1, j \mid i) > 0$, therefore M is irreducible.

Now, since the state space is finite, from Theorem 1.21, it follows that M has a unique stationary distribution. ■

We will also prove a lemma about the eigenvalues of a strictly positive Markov transition matrix M, which is similar to a symmetric PSD matrix.

Lemma 2.12. *Let $M \in Mat_n(\mathbb{R})$ be a strictly positive Markov transition matrix and let $M_s \in Mat_n(\mathbb{R})$ be a symmetric positive semi-definite matrix which is similar to M. Then the eigenvalues of M are*

$$1 = \lambda_0 > \lambda_1 \geqslant \ldots \geqslant \lambda_{n-1} \geqslant 0.$$

Proof. We will prove the lemma in three steps. First, we will show that all eigenvalues are non-negative. Secondly, we will show that no eigenvalue is greater than 1. And lastly, we will show that there exists a unique eigenvalue $\lambda_0 = 1$.

1. Since M_s is PSD then from Theorem 1.4 we know that all its eigenvalues are non-negative. From the similarity of M and M_s, it follows that the eigenvalues of M are also non-negative.

2. Let v be an eigenvector of M and $\lambda \geqslant 0$ the corresponding eigenvalue. Denote the index of the largest coordinate (by absolute value) by $i_{\max} = \operatorname{argmax}_i |v_i|$. Since M is a row-stochastic matrix, the following holds:

$$\begin{aligned} |\lambda|\,|v_{i_{\max}}| = |\lambda v_{i_{\max}}| = |(Mv)_{i_{\max}}| &= \left|\sum_{j=0}^{n-1} M_{i_{\max} j} v_j\right| \\ \leqslant \sum_{j=0}^{n-1} M_{i_{\max} j}\,|v_j| &\leqslant \sum_{j=0}^{n-1} M_{i_{\max} j}\,|v_{i_{\max}}| = |v_{i_{\max}}|\,, \end{aligned}$$

   hence $|\lambda| \leqslant 1$. Since $\lambda \geqslant 0$, we obtain $\lambda \leqslant 1$.

3. From Theorem 2.11 we know that M has a unique stationary distribution. From Theorem 1.21 we know that a left eigenvector of M is the stationary distribution with corresponding eigenvalue 1. From Theorem 1.22 we know that the eigenvalue $\lambda_0 = 1$ is algebraically simple.

Therefore we have shown that the eigenvalues of M are of form $1 = \lambda_0 > \lambda_1 \geqslant \ldots \geqslant \lambda_{n-1} \geqslant 0$. ■

Since M_s is a symmetric real matrix, we can eigen-decompose it, meaning

$$M_s = V \Lambda V^T$$

where $V \in \mathrm{Mat}_n(\mathbb{R})$ is a matrix with its columns being orthonormal eigenvectors of M_s and $\Lambda \in \mathrm{Mat}_n(\mathbb{R})$ is a diagonal matrix with eigenvalues $\{\lambda_j\}_{j=0}^{n-1}$ on its diagonal. From M_s being similar to M it follows that

$$M = D^{-1/2} M_s D^{1/2} = D^{-1/2} V \Lambda V^T D^{1/2} = \left(D^{-1/2} V\right) \Lambda \left(D^{1/2} V\right)^T = \Psi \Lambda \Phi^T.$$

Therefore

$$M^t = \Psi \Lambda^t \Phi^T.$$

Now, recall that $p(t, x_j \mid x_i)$ denotes the probability of travelling from state x_i to the state x_j in $t \in \mathbb{N}$ steps. By the Chapman-Kolmogorov equation (see [GS20, Theorem (7), §6.1]), $\left(M^t\right)_{ij} = p(t, x_j \mid x_i)$. Therefore

$$\begin{aligned}\left(M^t\right)_{ij} = p(t, x_j \mid x_i) &= \sum_{k=0}^{n-1} \lambda_k^t \psi_k(x_i)\varphi_k(x_j) \\ &= \psi_0(x_i)\varphi_0(x_j) + \sum_{k=1}^{n-1} \lambda_k^t \psi_k(x_i)\varphi_k(x_j),\end{aligned}$$

Now note that any constant vector $\boldsymbol{c} \in \mathbb{R}^{n\times 1}$ is a right eigenvector with eigenvalue 1. Indeed,

$$M\boldsymbol{c} = \left(\sum_{j=0}^{n-1} M_{0j}c, \ldots, \sum_{j=0}^{n-1} M_{(n-1)j}c\right)^T = \boldsymbol{c}.$$

Since M is irreducible, from Theorem 1.22 we know that the eigenvalue $\lambda_0 = 1$ is simple; therefore, ψ_0 is also a constant vector. Recalling that $\psi_0 = D^{-1/2}v_0$, this means $\psi_0(x_i) = D_{ii}^{-1/2}v_0(x_i)$ is the same for every $x_i \in \mathcal{X}$. It is worth noting that from ψ_0 being constant it follows that $v_0(x_i)$ is proportional to $D_{ii}^{1/2}$ for every $x_i \in \mathcal{X}$.

We can now define diffusion distance at time $t \in \mathbb{N} \cup \{0\}$ as

$$\begin{aligned}D_t^2(x_i, x_j) &\coloneqq \|p(t, x|x_i) - p(t, x|x_j)\|_w^2 \\ &\coloneqq \sum_{x\in\mathcal{X}} \left(p(t, x|x_i) - p(t, x|x_j)\right)^2 w(x),\end{aligned} \tag{1}$$

where we choose the weight $w(x) = 1/\varphi_0(x)$. Note that this weight is well-defined, because v_0 is a positive eigenvector and the diagonal matrix $D^{1/2}$ has positive entries on its diagonal. We can also denote the mapping between the original space and the first k non-trivial eigenvectors as the *diffusion map*

$$\Psi_t(x_i) \coloneqq \left(\lambda_1^t\psi_1(x_i), \ldots, \lambda_k^t\psi_k(x_i)\right), \tag{2}$$

with $\mathbb{N} \ni k \leqslant n-1$ (where $n = \operatorname{card}(\mathcal{X})$ is the number of vertices).

We now state the main theorem connecting diffusion distance to the Euclidean distance in the embedded space. Table 2 summarises the key objects used in the theorem.

Table 2: Key notation.

Symbol	Meaning	Note
$\mathcal{X}$	dataset of n points in $\mathbb{R}^p$	-
W	$\mathrm{Mat}_n(\mathbb{R}) \ni W,\ W_{ij} = k(x_i, x_j)$	Positive, symmetric, PSD
D	$\mathrm{Mat}_n(\mathbb{R}) \ni D,\ D_{ii} = \sum_{j=0}^{n-1} W_{ij}$	Diagonal
M	$D^{-1}W$	Row-stochastic
M_s	$D^{-1/2}WD^{-1/2}$	Symmetric, similar to M
v_j	j-th eigenvector of M_s	$\|v_j\|_2 = 1$
λ_j	j-th eigenvalue of M (and M_s)	-
$\psi_j = D^{-1/2}v_j$	j-th right eigenvector of M	-
$\varphi_j = D^{1/2}v_j$	j-th left eigenvector of M	-
$\Psi_t(x_i)$	Diffusion map of x_i at time t	-
$D_t^2(x_i, x_l)$	Diffusion distance between x_i and x_l	-

Theorem 2.13 (see [Nad+05, Eq. (11)]). *Let $\Psi_t : \mathcal{X} \to \mathbb{R}^{n-1}$ be a diffusion map defined using all $(n-1)$ non-trivial eigenvectors. The diffusion distance on $\mathcal{X}$ is equal, up to a scaling by a constant, to the Euclidean distance in $\mathbb{R}^{n-1}$.*

$$D_t^2(x_i, x_l) = \frac{1}{c}\sum_{j=1}^{n-1} \left(\lambda_j^t\right)^2 (\psi_j(x_i) - \psi_j(x_l))^2 = \frac{1}{c}\left\|\Psi_t(x_i) - \Psi_t(x_l)\right\|^2, \tag{3}$$

where $c \coloneqq \psi_0(x_i) = \dfrac{v_0(x_i)}{D_{ii}^{1/2}}$ which is independent of x_i, since ψ_0 is constant.

Proof. We start by rewriting the diffusion distance using the representation (1):

$$\begin{aligned}
D_t^2(x_i, x_l) &= \sum_{x\in\mathcal{X}} (p(t, x|x_i) - p(t, x|x_l))^2 w(x) \\
&= \sum_{x\in\mathcal{X}} \left(\sum_{j=0}^{n-1} \lambda_j^t(\psi_j(x_i) - \psi_j(x_l))\varphi_j(x)\right)^2 \frac{1}{\varphi_0(x)} \\
&= \sum_{x\in\mathcal{X}} \left(\sum_{j=0}^{n-1} \lambda_j^t(\psi_j(x_i) - \psi_j(x_l))\varphi_j(x)\right)\left(\sum_{k=0}^{n-1} \lambda_k^t(\psi_k(x_i) - \psi_k(x_l))\varphi_k(x)\right) \frac{1}{\varphi_0(x)} \\
&= \sum_{j,k=0}^{n-1} \lambda_j^t(\psi_j(x_i) - \psi_j(x_l))\lambda_k^t(\psi_k(x_i) - \psi_k(x_l)) \sum_{x\in\mathcal{X}} \frac{\varphi_j(x)\varphi_k(x)}{\varphi_0(x)}.
\end{aligned}$$

Before continuing, we will show that

$$\sum_{x\in\mathcal{X}} \frac{\varphi_j(x)\varphi_k(x)}{\varphi_0(x)} = \frac{1}{c}\delta_{jk}.$$

First recall that from the orthonormality of the eigenvectors v_j of matrix M_s and the forms $\varphi_j = D^{1/2}v_j$, $\psi_j = D^{-1/2}v_j$ we get $\langle\psi_j, \varphi_k\rangle = \delta_{jk}$. Thus it is sufficient to show that $\dfrac{\varphi_j(x)}{\varphi_0(x)} = \dfrac{1}{c}\psi_j(x)$ for all $x \in \mathcal{X}$, because then $\displaystyle\sum_{x\in\mathcal{X}} \frac{\varphi_j(x)\varphi_k(x)}{\varphi_0(x)} = \frac{1}{c}\sum_{x\in\mathcal{X}} \varphi_k(x)\psi_j(x) = \frac{1}{c}\langle\psi_j, \varphi_k\rangle$.

Now for any $j \in \{0, \dots, n-1\}$ and any x_i,

$$\begin{aligned}\varphi_j(x_i) &= \left(D^{1/2}v_j\right)(x_i) = \sqrt{D_{ii}}v_j(x_i),\\ \psi_j(x_i) &= \left(D^{-1/2}v_j\right)(x_i) = v_j(x_i)/\sqrt{D_{ii}}.\end{aligned}$$

Taking the ratio now

$$\begin{aligned}\frac{\varphi_j(x_i)}{\varphi_0(x_i)} &= \frac{\sqrt{D_{ii}}v_j(x_i)}{\sqrt{D_{ii}}v_0(x_i)} = \frac{1}{v_0(x_i)}v_j(x_i) = \frac{D_{ii}^{1/2}}{v_0(x_i)}\frac{v_j(x_i)}{D_{ii}^{1/2}}\\ &= \frac{D_{ii}^{1/2}}{v_0(x_i)}\psi_j(x_i) \overset{(*)}{=} \frac{1}{c}\psi_j(x_i) = \frac{1}{c}\psi_j(x_i),\end{aligned}$$

where $(*)$ holds because ψ_0 is a constant vector and therefore

$$c = \psi_0(x_i) = D_{ii}^{-1/2}v_0(x_i)$$

for any $i \in \{0, \dots, n-1\}$. This shows that

$$\sum_{x\in\mathcal{X}} \frac{\varphi_j(x)\varphi_k(x)}{\varphi_0(x)} = \frac{1}{c}\sum_{x\in\mathcal{X}} \psi_j(x)\varphi_k(x) = \frac{1}{c}\delta_{jk}.$$

Now we can continue with the proof of the theorem:

$$\begin{aligned}&\sum_{j,k=0}^{n-1} \lambda_j^t(\psi_j(x_i) - \psi_j(x_l))\lambda_k^t(\psi_k(x_i) - \psi_k(x_l)) \sum_{x\in\mathcal{X}} \frac{\varphi_j(x)\varphi_k(x)}{\varphi_0(x)}\\ &= \frac{1}{c}\sum_{j,k=0}^{n-1} \lambda_j^t(\psi_j(x_i) - \psi_j(x_l))\lambda_k^t(\psi_k(x_i) - \psi_k(x_l))\langle\psi_j, \varphi_k\rangle\\ &= \frac{1}{c}\sum_{j,k=0}^{n-1} \lambda_j^t(\psi_j(x_i) - \psi_j(x_l))\lambda_k^t(\psi_k(x_i) - \psi_k(x_l))\delta_{jk}\\ &= \frac{1}{c}\sum_{j=0}^{n-1} \left(\lambda_j^t\right)^2(\psi_j(x_i) - \psi_j(x_l))^2.\end{aligned}$$

Now, since $\lambda_0 = 1$, ψ_0 is a constant vector, we get $(\psi_0(x_i) - \psi_0(x_l)) = 0$, thus

$$\frac{1}{c}\sum_{j=0}^{n-1}\left(\lambda_j^t\right)^2\left(\psi_j(x_i)-\psi_j(x_l)\right)^2 = \frac{1}{c}\sum_{j=1}^{n-1}\left(\lambda_j^t\right)^2\left(\psi_j(x_i)-\psi_j(x_l)\right)^2$$
$$= \frac{1}{c}\left\|\Psi_t(x_i)-\Psi_t(x_l)\right\|^2.$$

That is exactly what we wanted to show. Note that the choice of the weight $w(x) = 1/\varphi_0(x)$ is essential for the theorem to hold. ■

3 Applications

As mentioned in the introduction, diffusion maps have many applications, including in computer vision and in statistical data analysis. In computer vision, we are often interested in determining whether two shapes are intrinsically similar, regardless of their positions in the ambient space. In data analysis, we often work with high-dimensional data and are interested in whether it can be represented in a computationally cheaper, lower-dimensional space without losing important information.

In this section, we illustrate the applications of diffusion maps on three artificial datasets and one well-established dataset. The first example is a dumbbell-shaped dataset, intended to give an intuitive sense of the relationship between diffusion distance on the graph and the Euclidean distance in the embedded space. We follow this up by applying diffusion maps to a dataset shaped like a toroidal helix, demonstrating how diffusion maps can be used for dimensionality reduction to retain the intrinsic geometry, and how a classical linear method, principal component analysis (see [Jol02, Section 1.1]), fails to do so. For the third artificial example, we show how diffusion maps can be used to solve correspondence problems in a controlled environment using two intrinsically identical tori. This naturally leads to our last example, which shows how diffusion maps can be applied to intrinsically identical real-world data, and identifies their limitations. For this, we chose high-resolution human figures from the Dynamic FAUST dataset (see [Bog+17]). All code for generating the datasets, applying diffusion maps and calculating the errors is available on GitHub (see [Nur26]).

3.1 Artificial Examples

In all three of the following examples, we will illustrate the embeddings in either two- or three-dimensional space. This deviates from Theorem 2.13, which states that diffusion distance is equivalent to the Euclidean distance in the embedded $(n-1)$-dimensional space. Nonetheless, the embeddings capture the essence well and are therefore valuable.

Before starting with the specific examples, we provide the parameters for all three examples in Table 3. For all datasets, the Gaussian kernel (see Example 2.8) was used to construct a weighted graph from the dataset.

Table 3: Parameters used for the three examples in Section 3.1.

Dataset	n **(data points)**	ε **(width)**[2]	**Data dim.**	**Embedding dim.**
Dumbbell	515	0.05	2	1
Toroidal helix	3000	0.05	3	2
Tori	5000	0.05	3	2

[2] ε is the width parameter of the Gaussian kernel (see Example 2.8).

3.1.1 Dumbbell-shaped dataset

To illustrate what diffusion maps capture, we chose to embed a dumbbell-shaped dataset (for exact parameters see Table 3). The dataset in Figure 2 contains two clusters (coloured turquoise and grey) connected by a narrow neck (coloured light green). We highlight three points: points A and B are chosen to be near the bottleneck in the grey and turquoise cluster, respectively. Point C is chosen to be farther from the "bridge" in the turquoise cluster.

Figure 2 shows the Euclidean distance between pairs A, B and B, C. With respect to the Euclidean distance, points A and B are closer to each other than points B and C.

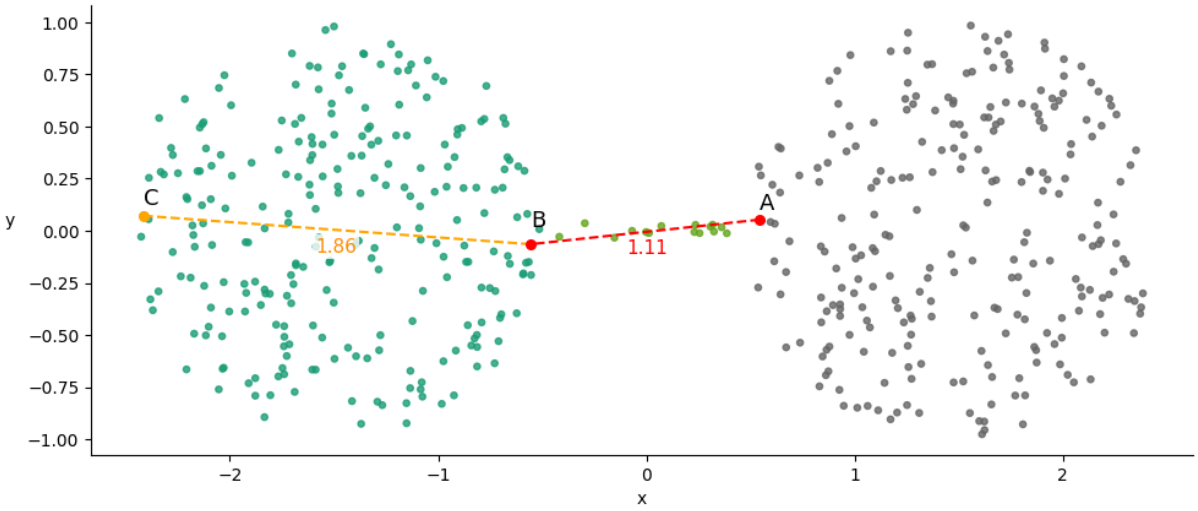


Figure 2: Dumbbell-shaped dataset, with highlighted points and the Euclidean distances between them.

However, if we now look at the weighted paths connecting the respective pairs (see Figure 3), it is clear that points B and C are much better connected than points A and B. This is expected since B and C lie in the same cluster. Even though the graph is fully connected, the paths are weighted using the Gaussian kernel (see Example 2.8), so points within the cluster are connected via edges with larger weights. This means that the random walk is more likely to stay within the cluster it starts from, and hence the diffusion distance between points B and C is smaller than between points A and B.

Note that in Figure 3 the number of simulated paths was bounded by 2500 and the number of steps in each path was bounded by 45. Only paths that reached the desired endpoint were visualised (for example, for paths from B to C, only paths that reached point C in no more than 45 steps are shown). Additionally, the graph in Figure 3 is fully connected, but since the width of the edges depends on their weight, edges with very small weights are not visible in the plot.

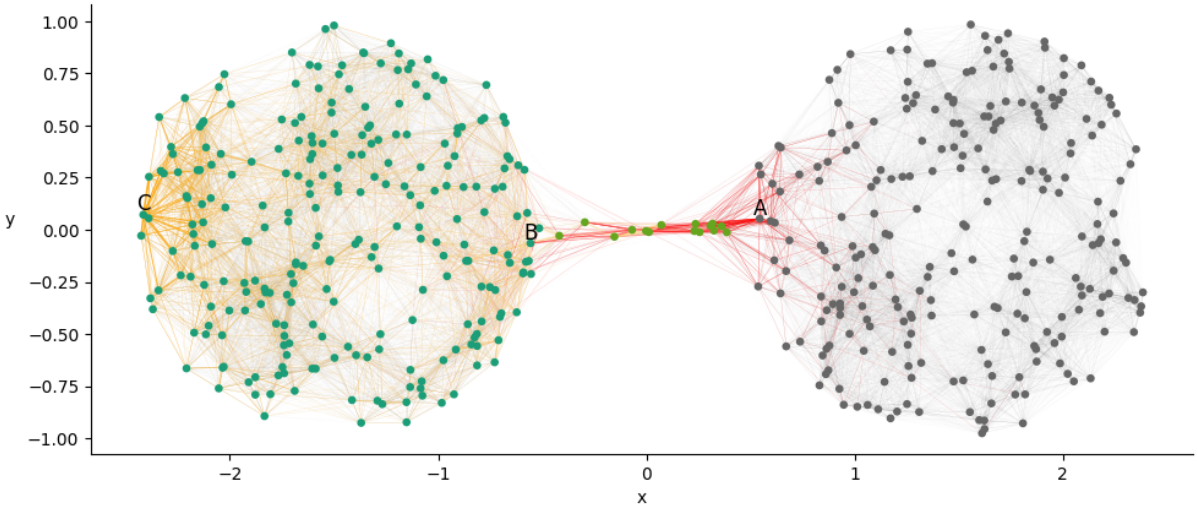


Figure 3: Dumbbell-shaped dataset, with red paths connecting points A and B and orange paths connecting points B and C.

After embedding the graph following the construction stated in Subsection 2.2 but using only the first non-trivial eigenvector and eigenvalue, the resulting embedding lies in $\mathbb{R}$. Figure 4 shows the resulting embedding in $\mathbb{R}$. As expected, points B and C, which have a smaller diffusion distance, are mapped closer in $\mathbb{R}$ than points A and B, which have a larger diffusion distance. The two clusters are cleanly separated in the embedding, with the bottleneck points being mapped between them.

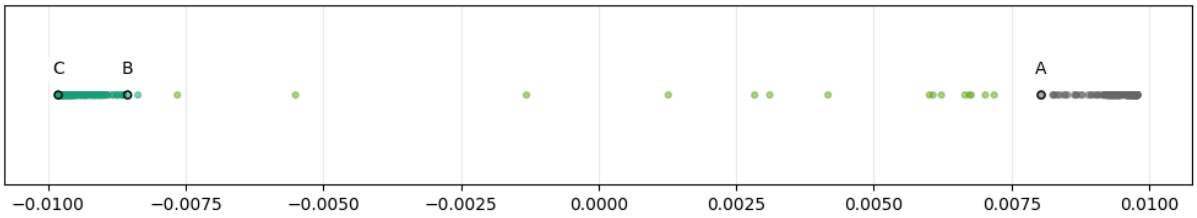


Figure 4: Dumbbell-shaped graph mapped to $\mathbb{R}$ with highlighted points A, B, C.

3.1.2 Toroidal-helix-shaped dataset

To illustrate which points on the toroidal helix are close to one another with respect to diffusion distance, we applied a colour gradient to the generated dataset. The gradient parametrises positions along the toroidal helix, meaning points which are intrinsically close are similar in colour. This allows us later to illustrate how diffusion maps capture the intrinsic geometry that principal component analysis (PCA, for short) does not. We will not discuss in detail why using diffusion maps for dimensionality reduction is justified, since it is outside the scope of this thesis. In short, the right eigenvectors of the Markov matrix form a basis of $\mathbb{R}^n$, but often the first $k \leqslant n-1$ non-trivial eigenvectors capture the most prominent features with respect to diffusion, and therefore it is sufficient to embed the data into $\mathbb{R}^k$ (see [Nad+05, Section 2]). On the other hand, PCA projects the data onto $l \leqslant \min(p, n-1)$ eigenvectors of a covariance matrix (where n is the number of data

points, and p is the dimensionality of data points), and thus captures the directions of greatest variance. Figure 5 illustrates the coloured dataset resembling a toroidal helix in $\mathbb{R}^3$.

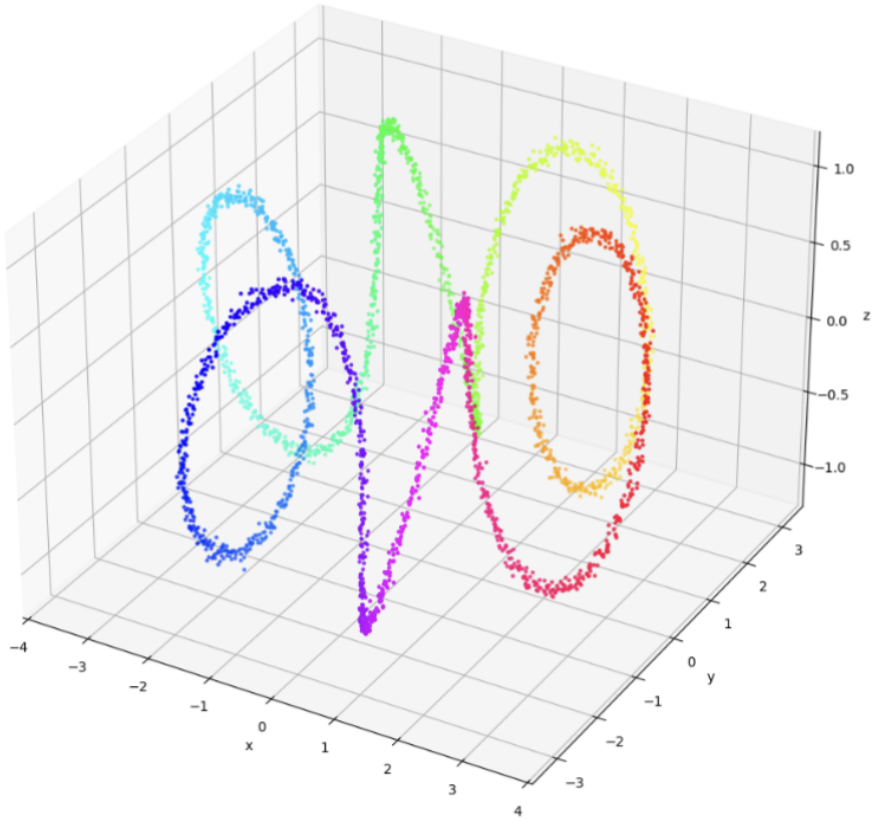
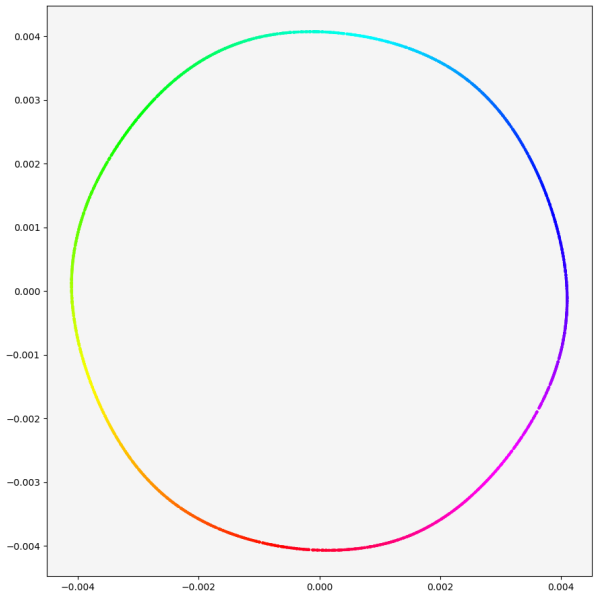


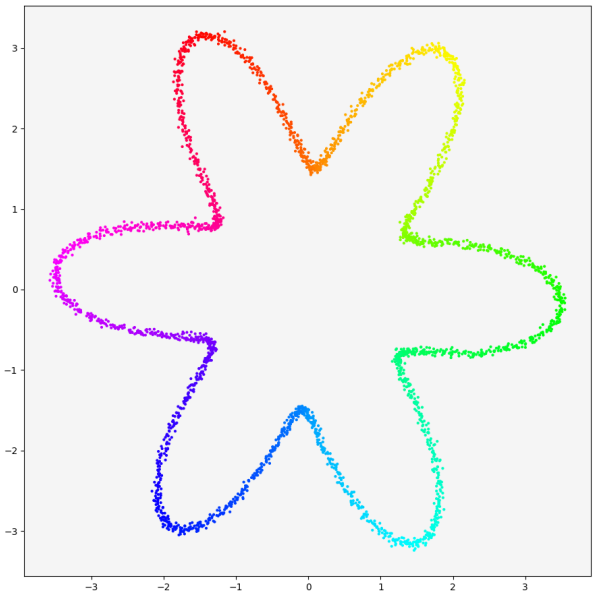
Figure 5: Toroidal-helix-shaped dataset in $\mathbb{R}^3$.

After applying diffusion maps to the shape and mapping it into $\mathbb{R}^2$, using two eigenvectors corresponding to the two largest non-trivial eigenvalues, we end up with a circular shape shown in Figure 6a.

This coincides nicely with what is expected, since points which are farthest from each other on the toroidal helix with respect to diffusion distance are also far apart from each other in the embedded space with respect to the Euclidean distance. Comparing the result to applying PCA on the same dataset and embedding it to $\mathbb{R}^2$ (Figure 6b), we can clearly see that the intrinsic geometry is not captured.



(a) Diffusion maps for dimensionality reduction on the toroidal helix.



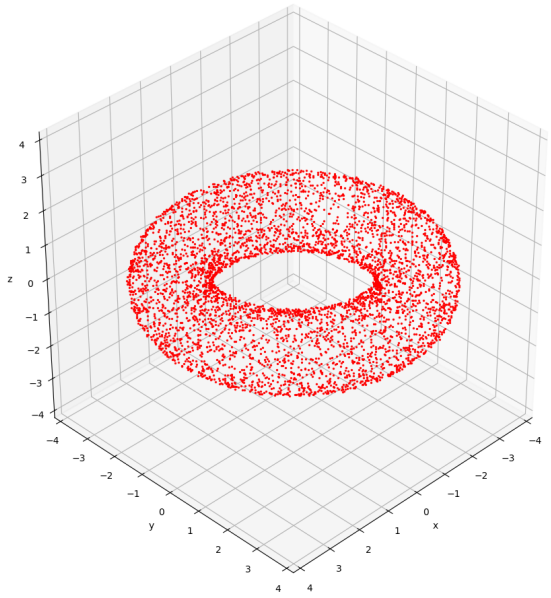
(b) PCA for dimensionality reduction on the toroidal helix.

Figure 6: Comparison of results of applying diffusion maps and PCA for dimensionality reduction.

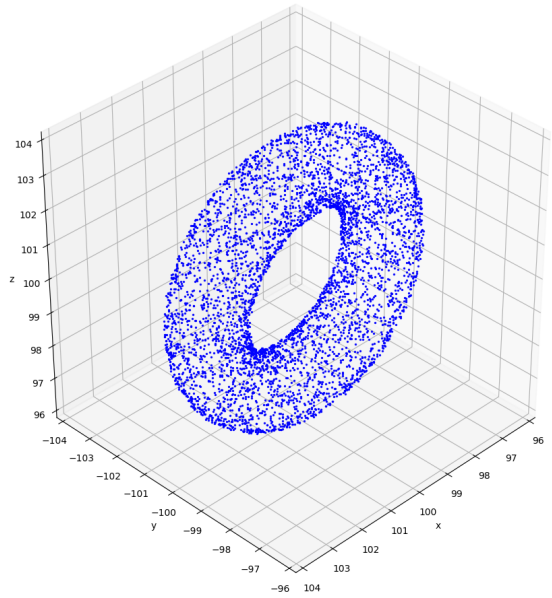
3.1.3 Torus-shaped dataset

As mentioned, in computer vision, diffusion maps can be used to determine if two shapes are intrinsically similar. This kind of problem is called a correspondence problem. Formally, correspondence problems are tasks where we have $N \in \mathbb{N}$ graphs $G_1 = (V_1, E_1), \ldots, G_N = (V_N, E_N)$ and we are interested in finding subsets of vertices $V_2' \subset V_2, \ldots, V_k' \subset V_k$ $(\mathbb{N} \ni k \leqslant N)$ which are most similar (with respect to some metric) to a given subset $V_1' \subset V_1$.

For this example, we chose to work with two tori. We will first provide a visual solution to the task by comparing embeddings in $\mathbb{R}^2$, and then confirm the result numerically. We generated the first torus (in red) to be centred around the origin $(0, 0, 0)$ in $\mathbb{R}^3$ (Figure 7a). The second one is an identical copy of the first, shifted to be centred around $(100, 100, 100)$ and rotated 90 degrees around the x-axis (Figure 7b). This is a good example of a controlled geometric setting, since the Markov matrices, and hence the embeddings, are identical because the Gaussian kernel is invariant under shifts and rotations.



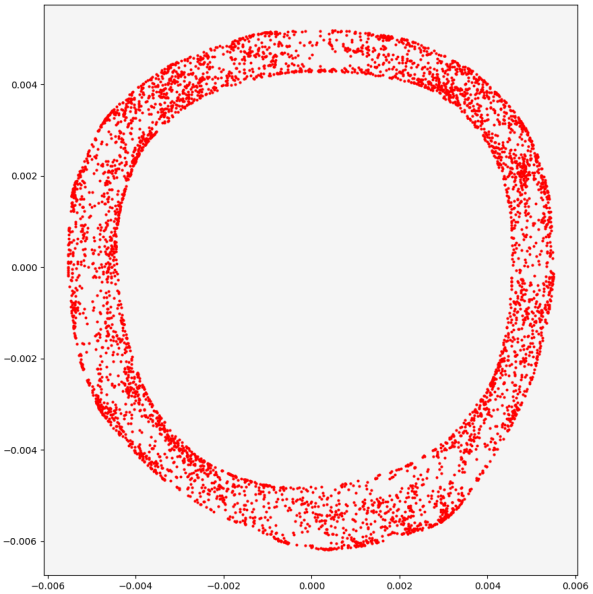
(a) Original torus-shape dataset in $\mathbb{R}^3$, denoted X_1.



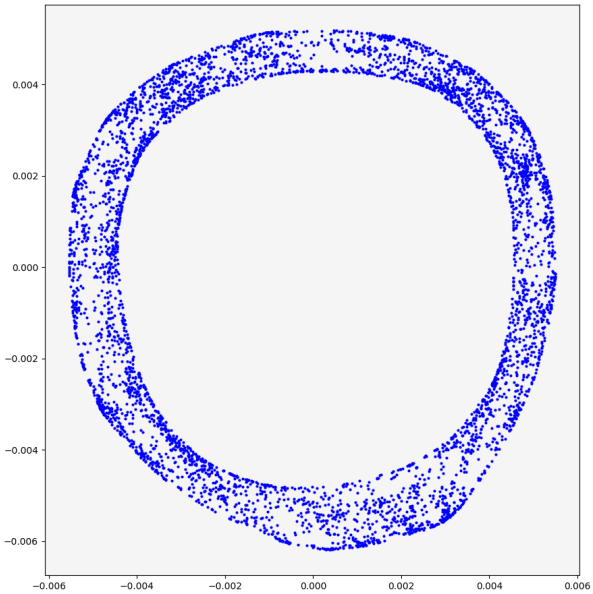
(b) Transformed torus-shape dataset in $\mathbb{R}^3$, denoted X_2.

Figure 7: The original and transformed tori in $\mathbb{R}^3$.

After embedding both tori into $\mathbb{R}^2$ separately and comparing the mappings, it is easy to see that the embeddings are identical (Figures 8a and 8b).



(a) Original torus embedded, denoted $\Psi_{X_1}(X_1)$.



(b) Transformed torus embedded, denoted $\Psi_{X_2}(X_2)$.

Figure 8: Original and transformed tori embedded into $\mathbb{R}^2$ using diffusion maps.

The identical embeddings in Figure 8 confirm visually that the diffusion maps captured the intrinsic geometry. To confirm our result numerically, we calculated the error $r = \sum_{i=1}^{\text{card}(X)} \|X_i - Y_i\|_2$ (where

X, Y are datasets of the same cardinality) between the original datasets X_1, X_2 and between the embeddings $\Psi_{X_1}(X_1), \Psi_{X_2}(X_2)$:

$$r_{\text{original}} = \sum_{i=1}^{\text{card}(X_1)} \|(X_1)_i - (X_2)_i\|_2 \approx 866216.966,$$
$$r_{\text{embedded}} = \sum_{i=1}^{\text{card}(\Psi_{X_1}(X_1))} \|\Psi_{X_1}(X_1)_i - \Psi_{X_2}(X_2)_i\|_2 \approx 3.612 \cdot 10^{-7}.$$

Based on Theorem 2.13, we would expect the sum of pairwise distances between the embedded shapes r_{embedded} to be exactly zero (assuming constant sign choices for the eigenvectors). The computed value is close, and the difference is most likely caused by numerical errors.

3.2 Diffusion Maps on a Non-Rigid Correspondence Problem

In the example of two tori from Subsection 3.1.3, the geometry of the data was controlled: the second torus was a shifted and rotated copy of the original one, and therefore their weight matrices were identical. For real-world correspondence problems, such good conditions are rare. In this section, we will illustrate how diffusion maps perform for solving more complex correspondence problems.

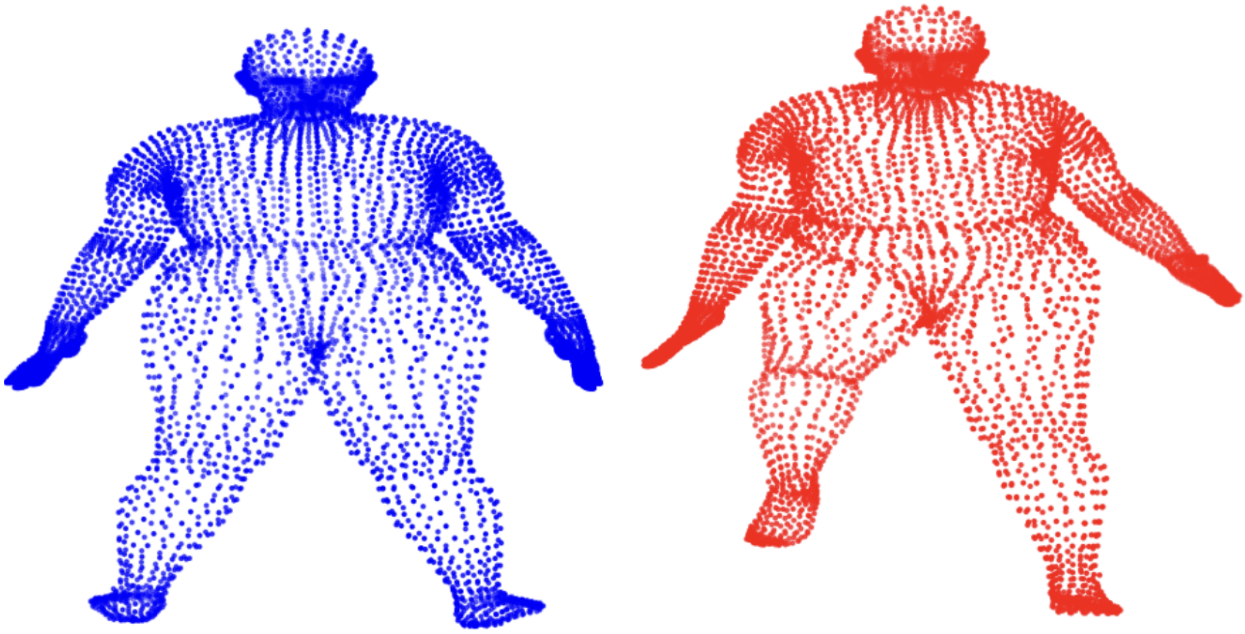
We chose to work with two figures from the Dynamic FAUST dataset (see [Bog+17]), which contains high-resolution human-shaped figures in different poses. This is a more difficult task than the one with tori, since even though the human figures in different poses are intrinsically identical, the change in pose is a non-rigid transformation, which causes the weight matrices to differ. Indeed, when a person moves naturally (running, for example), the Euclidean distance between body parts changes. Since the weights are built using the Gaussian kernel (see Example 2.8), which is influenced by the ambient space, the weight matrices corresponding to different poses do differ. Hence, the embeddings are not identical.

In some cases, the Gaussian kernel parameters can be tuned to better capture the prominent features. Exploring how to choose the optimal parameters is outside the scope of this thesis, but briefly, for a small ε, only local features are captured, and as ε approaches ∞, the only thing that diffusion maps capture is the number of data points. We chose $\varepsilon = 0.5$ based on a few trials. Table 4 summarises the parameters of this example.

Table 4: Parameters used for the FAUST example in Section 3.2.

Dataset	n **(data points)**	ε **(width)**^{2}	**Data dim.**	**Embedding dim.**
DFAUST	6890	0.5	3	3

Figure 9 shows two human-shaped figures in the original space $\mathbb{R}^3$. In the first figure (Figure 9a) the person is standing still while in the second figure (Figure 9b) the person is running-in-place.



(a) Dataset X in the shape of a standing figure.

(b) Dataset Y in the shape of a running-in-place figure.

Figure 9: Two figures from the Dynamic FAUST dataset in the space $\mathbb{R}^3$.

Figure 10 shows the embeddings in $\mathbb{R}^3$. We can see that they do not coincide perfectly but are relatively similar.

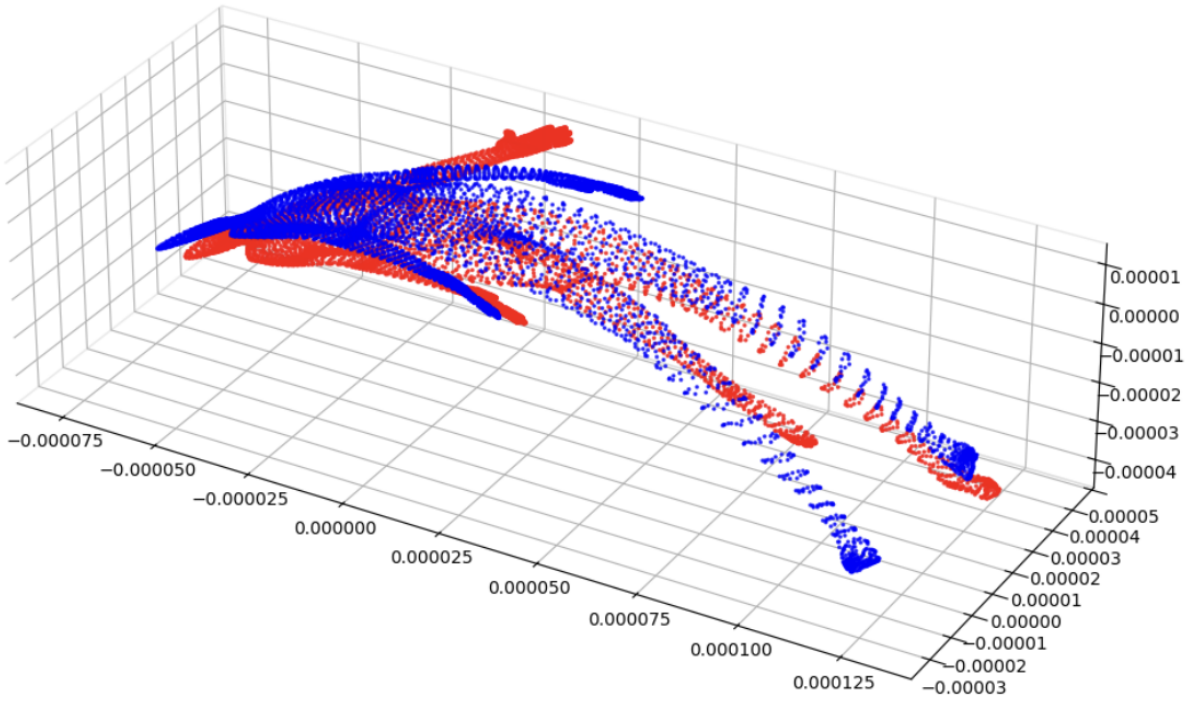


Figure 10: Dynamic FAUST figures embedded into $\mathbb{R}^3$ using diffusion maps.

Now confirming the results numerically:

$$r_{\text{original}} = \sum_{i=1}^{\text{card}(X)} \|X_i - Y_i\|_2 \approx 1078.315,$$
$$r_{\text{embedded}} = \sum_{i=1}^{\text{card}(\Psi_X(X))} \|\Psi_X(X_i) - \Psi_Y(Y_i)\|_2 \approx 0.0896.$$

Therefore, we have illustrated that even for solving correspondence problems with non-rigidly transformed data, diffusion maps perform relatively well. Finally, it is worth noting again that performance could likely be improved by optimising kernel parameters (or possibly by choosing a different kernel altogether).

Summary

The main objectives of this thesis were to provide a self-contained proof of Theorem 2.13, identify sufficient conditions on the kernel function for the theorem to hold, and illustrate how diffusion maps perform on both synthetic and real-world data.

In Definitions and Preliminaries (see Section 1), we recalled necessary definitions and results from linear algebra and probability theory.

In Diffusion Distance and Diffusion Maps (see Section 2), we first proved the spectral theorem, which states that every real symmetric matrix has real eigenvalues and its eigenvectors form an orthonormal basis of $\mathbb{R}^n$. We then developed the theory of diffusion maps in a self-contained way. We stated explicit sufficient conditions on the kernel (strict positivity, symmetry, and positive semi-definiteness) and proved Theorem 2.13, which states that the diffusion distance is equivalent to the Euclidean distance in the embedded space. The original paper by Nadler et al. [Nad+05] proves the theorem using the Gaussian kernel but does not explicitly state the conditions under which the result holds for general kernels and relies on auxiliary results that are not always stated.

In Applications (see Section 3), we applied the theory to three synthetic datasets and a real-world dataset. The synthetic examples demonstrated how diffusion maps work at an intuitive level (Subsection 3.1.1); how they can be used for dimensionality reduction, including how they differ from PCA, a classical linear method (Subsection 3.1.2); and, lastly, how they can be used to solve correspondence problems (Subsection 3.1.3). We concluded the section by applying diffusion maps to two figures from the Dynamic FAUST dataset. Since the Gaussian kernel is not invariant under non-rigid transformation, the embeddings did not align as cleanly as in the synthetic examples. However, when embedding the figures into $\mathbb{R}^3$, the results were similar. Checking the correspondence numerically confirmed that diffusion maps also perform well with complex real-world data.

There are several directions in which this work could be extended. We identified sufficient conditions on the kernel, but we believe they can be weakened: strict positivity is used to guarantee irreducibility and aperiodicity, positive semi-definiteness to guarantee non-negative eigenvalues. We believe weaker conditions that ensure the same properties would suffice. Identifying exact necessary and sufficient conditions would be the natural next step. Furthermore, the Dynamic FAUST example (Subsection 3.2) raises the question of which kinds of kernels are better suited for solving non-rigid correspondence problems. Finally, providing rigorous mathematical justifications for the use of diffusion maps for dimensionality reduction would be a valuable direction for future work. In particular, formalising the spectral-gap argument, which is one of the key criteria for determining the minimal number of dimensions to keep.

References

[Ant26] Anthropic. *Claude Opus 4.7*. 2026. URL: https://claude.ai/.

[Axl20] Sheldon Axler. *Measure, integration & real analysis*. English. Vol. 282. Grad. Texts Math. Cham: Springer, 2020. ISBN: 978-3-030-33142-9; 978-3-030-33143-6. DOI: 10.1007/978-3-030-33143-6.

[Axl24] Sheldon Axler. *Linear algebra done right*. English. 4th edition. Undergraduate Texts Math. Cham: Springer, 2024. ISBN: 978-3-031-41025-3; 978-3-031-41026-0. DOI: 10.1007/978-3-031-41026-0.

[Bog+17] Federica Bogo et al. "Dynamic FAUST: Registering Human Bodies in Motion". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[Goo26] Google. *Gemini 3*. 2026. URL: https://gemini.google.com/.

[GS20] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and random processes*. English. 4th edition. Oxford: Oxford University Press, 2020. ISBN: 978-0-19-884760-1; 978-0-19-884759-5; 978-0-19-884762-5.

[HJ13] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. English. 2nd ed. Cambridge: Cambridge University Press, 2013. ISBN: 978-0-521-54823-6; 978-0-521-83940-2.

[Jol02] I. T. Jolliffe. *Principal component analysis*. English. 2nd ed. Springer Ser. Stat. New York, NY: Springer, 2002. ISBN: 0-387-95442-2. DOI: 10.1007/b98835.

[Laa23] Valdis Laan. *Algebra I*. University of Tartu lecture notes (MTMM.00.038). 2023.

[Nad+05] Boaz Nadler et al. "Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators". In: *Advances in Neural Information Processing Systems*. Vol. 18. 2005, pp. 955–962.

[Nur26] Riki-Taavi Nurm. *Code-For-BSc-Thesis*. https://github.com/ri74ki74/Diffusion-Distance-and-Diffusion-Maps-Theory-and-Applications/tree/main. Accessed: 27.05.2026. 2026.

[Ope25] OpenAI. *ChatGPT (GPT-5)*. 2025. URL: https://chat.openai.com/.

[Str16] Gilbert Strang. *Introduction to linear algebra*. English. 5th edition. Wellesley, MA: Wellesley-Cambridge Press, 2016. ISBN: 978-0-9802327-7-6. URL: math.mit.edu/~gs/linearalgebra/.

[VW25] Gregory Valiant and Mary Wootters. *Lecture #14: The Fundamental Theorem of Markov Chains and Stationary Distributions*. Stanford University CS265/CME309 Course Notes. 2025. URL: https://web.stanford.edu/class/cs265/Lectures/Lecture14/l14.pdf (visited on 05/19/2026).

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Riki-Taavi Nurm,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Diffusion Distance and Diffusion Maps: Theory and Applications", mille juhendajad on professor Raul Vicente Zafra ja kaasprofessor Johann Langemets, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Riki-Taavi Nurm
27.05.2026