THESIS

MULTIDIMENSIONAL SCALING: INFINITE METRIC MEASURE SPACES

Submitted by

Lara Kassab

Department of Mathematics

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2019

Master's Committee:

Advisor: Henry Adams

Michael Kirby
Bailey Fosdick

ABSTRACT

MULTIDIMENSIONAL SCALING: INFINITE METRIC MEASURE SPACES

Multidimensional scaling (MDS) is a popular technique for mapping a finite metric space into a low-dimensional Euclidean space in a way that best preserves pairwise distances. We study a notion of MDS on infinite metric measure spaces, along with its optimality properties and goodness of fit. This allows us to study the MDS embeddings of the geodesic circle S^1 into $\mathbb{R}^m$ for all m, and to ask questions about the MDS embeddings of the geodesic n-spheres S^n into $\mathbb{R}^m$. Furthermore, we address questions on convergence of MDS. For instance, if a sequence of metric measure spaces converges to a fixed metric measure space X, then in what sense do the MDS embeddings of these spaces converge to the MDS embedding of X? Convergence is understood when each metric space in the sequence has the same finite number of points, or when each metric space has a finite number of points tending to infinity. We are also interested in notions of convergence when each metric space in the sequence has an arbitrary (possibly infinite) number of points.

ACKNOWLEDGEMENTS

We would like to thank Henry Adams, Mark Blumstein, Bailey Fosdick, Michael Kirby, Henry Kvinge, Facundo Mémoli, Louis Scharf, the students in Michael Kirby's Spring 2018 class, and the Pattern Analysis Laboratory at Colorado State University for their helpful conversations and support throughout this project.

DEDICATION

I would like to dedicate this thesis to my parents.

TABLE OF CONTENTS

Chapter 1

Introduction

Multidimensional scaling (MDS) is a set of statistical techniques concerned with the problem of constructing a configuration of n points in a Euclidean space using information about the dissimilarities between the n objects. The dissimilarities need not be based on Euclidean distances; they can represent many types of dissimilarities between objects. The goal of MDS is to map the objects $x_1, \ldots, x_n$ to a configuration (or embedding) of points $f(x_1), \ldots, f(x_n)$ in $\mathbb{R}^m$ in such a way that the given dissimilarities d_{ij} are well-approximated by the Euclidean distances between $f(x_i)$ and $f(x_j)$. The different notions of approximation give rise to the different types of MDS, and the choice of the embedding dimension m is arbitrary in principle, but low in practice ($m = 1, 2,$ or 3).

MDS is an established multivariate analysis technique used in a multitude of disciplines. It mainly serves as a visualization technique for proximity data, the input of MDS, which is usually represented in the form of an $n \times n$ dissimilarity matrix. Proximity refers to similarity and dissimilarity measures; these measures are essential to solve many pattern recognition problems such as classification and clustering. A frequent source of dissimilarities is distances between high-dimensional objects, and in this case, MDS acts as an (often nonlinear) dimension reduction technique.

MDS is indeed an optimization problem because a perfect Euclidean embedding preserving the dissimilarity measures does not always exist. If the dissimilarity matrix can be realized exactly as the distance matrix of some set of points in $\mathbb{R}^m$ (i.e. if the dissimilarity matrix is *Euclidean*), then MDS will find such a realization. Furthermore, MDS can be used to identify the minimum such Euclidean dimension m admitting an isometric embedding. However, some dissimilarity matrices or metric spaces are inherently non-Euclidean (cannot be embedded into $\mathbb{R}^m$ for any m). When

a dissimilarity matrix is not Euclidean, then MDS produces a mapping into $\mathbb{R}^m$ that distorts the interpoint pairwise distances as little as possible, in a sense that can be made precise.

The various types of MDS arise mostly from the different loss functions they minimize, and they mainly fall into two categories: metric and non-metric MDS. A brief overview on the different types of MDS is given in Section 4.2. One of the main methods of MDS is commonly known as classical multidimensional scaling (cMDS), which minimizes a form of a loss function known as Strain. The classical MDS algorithm is algebraic and not iterative. Therefore, it is simple to implement, and is guaranteed to discover the optimal configuration in $\mathbb{R}^m$. In Section 4.2.1, we describe the algorithm, and discuss its optimality properties and goodness of fit.

A *metric measure space* is a triple (X, d_X, μ_X) where (X, d_X) is a compact metric space, and μ_X is a Borel probability measure on X. In this work, we study a notion of MDS on infinite metric measure spaces, which can be simply thought of as spaces of infinitely many points equipped with some probability measure. Our motivation is to prove convergence properties of MDS of metric measure spaces. That is, if a sequence of metric measure spaces X_n converges to a fixed metric measure space X as $n \to \infty$, then in what sense do the MDS embeddings of these spaces converge to the MDS embedding of X? Convergence is well-understood when each metric space has the same finite number of points, and also fairly well-understood when each metric space has a finite number of points tending to infinity. An important example is the behavior of MDS as one samples more and more points from a dataset. We are also interested in convergence when the metric measure spaces in the sequence perhaps have an infinite number of points. In order to prove such results, we first need to define the MDS embedding of an infinite metric measure space X, and study its optimal properties and goodness of fit.

In Section 6.3, we explain how MDS generalizes to possibly infinite metric measure spaces. We describe an infinite analogue to the classical MDS algorithm. Furthermore, in Theorem 6.4.3 we show that this analogue minimizes a Strain function similar to the Strain function of classi-

cal MDS. This theorem generalizes [3, Theorem 14.4.2], or equivalently [37, Theorem 2], to the infinite case. Our proof is organized analogously to the argument in [37, Theorem 2].

As a motivating example, we consider the MDS embeddings of the circle equipped with the (non-Euclidean) geodesic metric. By using the properties of circulant matrices, we carefully identify the MDS embeddings of evenly-spaced points from the geodesic circle into $\mathbb{R}^m$, for all m. As the number of points tends to infinity, these embeddings lie along the curve

$$\sqrt{2}\left(\cos\theta, \sin\theta, \tfrac{1}{3}\cos 3\theta, \tfrac{1}{3}\sin 3\theta, \tfrac{1}{5}\cos 5\theta, \tfrac{1}{5}\sin 5\theta, \ldots\right) \in \mathbb{R}^m.$$

Furthermore, we address convergence questions for MDS. Indeed, convergence is well-understood when each metric space has the same finite number of points [31], but we are also interested in convergence when the number of points varies and is possibly infinite. We survey Sibson's perturbation analysis [31] for MDS on a fixed number of n points. We survey results of [2, 16] on the convergence of MDS when n points $\{x_1, \ldots, x_n\}$ are sampled from a metric space according to a probability measure μ, in the limit as $n \to \infty$. We reprove these results under the (simpler) deterministic setting when points are not randomly chosen, and instead we assume that the corresponding finite measures $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ (determined by n points) converge to μ. This allows us, in Section 8.4, to consider the more general setting where we have convergence of *arbitrary* probability measures $\mu_n \to \mu$, where now each measure μ_n is allowed to have infinite support.

In Chapter 2, we survey related work. In Chapter 3, we present background information on proximities, metric spaces, spaces with various structures such as inner products or norms, and metric measure spaces. We present an overview on the theory of MDS in Chapter 4. We briefly describe the different types of MDS with most emphasis on classical MDS. In Chapter 5, we present necessary background information on operator theory and infinite-dimensional linear algebra. We define a notion of MDS for infinite metric measure spaces in Chapter 6. In Chapter 7, we identify

the MDS embeddings of the geodesic circle into $\mathbb{R}^m$, for all m, as a motivating example. Lastly, in Chapter 8, we describe different notions of convergence of MDS.

Chapter 2

Related Work

The reader is referred to the introduction of [38] and to [12, 14] for some aspects of history of MDS. Furthermore, the reader is referred to [33] for the theory of linear equations, mainly of the second kind, associated with the names of Volterra, Fredholm, Hilbert and Schmidt. The treatment has been modernised by the systematic use of the Lebesgue integral, which considerably widens the range of applicability of the theory. Among other things, this book considers singular functions and singular values as well.

There are a variety of papers that study some notion of robustness or convergence of MDS. In a series of papers [30–32], Sibson and his collaborators consider the robustness of multidimensional scaling with respect to perturbations of the underlying distance or dissimilarity matrix. Indeed, convergence is well-understood when each metric space has the same finite number of points [31], but we are also interested in convergence when the number of points varies and is possibly infinite. The paper [2] studies the convergence of MDS when more and more points are sampled independent and identically distributed (i.i.d.) from an unknown probability measure μ on X. The paper [16] presents a key result on convergence of eigenvalues of operators. Furthermore, [23, Section 3.3] considers embedding new points in psuedo-Euclidean spaces, [13, Section 3] considers infinite MDS in the case where the underlying space is an interval (equipped with some metric), and [9, Section 6.3] discusses MDS on large numbers of objects.

Some popular non-linear dimensionality reduction techniques besides MDS include Isomap [35], Laplacian eigenmaps [1], Locally Linear Embedding (LLE) [24], and Nonlinear PCA [27]. See also the recent paper [19] for a GPU-oriented dimensionality reduction algorithm that is inspired by Whitney's embedding theorem. The paper [41] makes a connection between kernel PCA and

metric MDS, remarking that kernel PCA is a form of MDS when the kernel is isotropic. The reader is referred to [2] for further relations between spectral embedding methods and kernel PCA.

2.1 Applications of MDS

MDS is an established multivariate analysis technique used in a multitude of disciplines like social sciences, behavioral sciences, political sciences, marketing, etc. One of the main advantages of MDS is that we can analyze any kind of proximity data, i.e. dissimilarity or similarity measures. For instance, MDS acts as an (often nonlinear) dimension reduction technique when the dissimilarities are distances between high-dimensional objects. Furthermore, when the dissimilarities are shortest-path distances in a graph, MDS acts as a graph layout technique [9]. Furthermore, MDS is used in machine learning in solving classification problems. Some related developments in machine learning include Isomap [35] and kernel PCA [26]. The reader is referred to [8–10] for further descriptions on various applications of MDS.

2.2 Multivariate Methods Related to MDS

There exists several multivariate methods related to MDS. Some include Principal Component Analysis (PCA), Correspondence Analysis, Cluster Analysis and Factor Analysis [8]. For instance, PCA finds a low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional input space. Classical MDS finds an embedding that preserves the interpoint distances, and it is equivalent to PCA when those distances are Euclidean. The reader is referred to [8, Chapter 24] for a detailed comparison between the first three methods and MDS. In Factor Analysis, the similarities between objects are expressed in the correlation matrix. With MDS, one can analyze any kind of similarity or dissimilarity matrix, in addition to correlation matrices.

Chapter 3

Preliminaries

We describe preliminary material on proximities and metrics, on inner product and normed spaces, on metric measure spaces, and on distances between metric measure spaces.

3.1 Proximities and Metrics

Proximity means nearness between two objects in a certain space. It refers to similarity and dissimilarity measures, which are essential to solve many pattern recognition problems such as classification and clustering. Two frequent sources of dissimilarities are high-dimensional data and graphs [9]. Proximity data between n objects is usually represented in the form of an $n \times n$ matrix. When the proximity measure is a *similarity*, it is a measure of how similar two objects are, and when it is a *dissimilarity*, it is a measure of how dissimilar two objects are. See [10, Section 1.3] for a list of some of the commonly used similarity and dissimilarity measures. MDS is a popular technique used in order to visualize proximities between a collection of objects. In this section, we introduce some of the terminology and define some of the proximity measures discussed in the following chapters.

Definition 3.1.1. An $(n \times n)$ matrix $\mathbf{D}$ is called a *dissimilarity matrix* if it is symmetric and

$$d_{rr} = 0, \quad \text{with} \quad d_{rs} \geq 0 \quad \text{for} \quad r \neq s.$$

The first property above is called refectivity, and the second property is called nonnegativity. Note that there is no need to satisfy the triangle inequality.

Definition 3.1.2. A function $d\colon X \times X \to \mathbb{R}$ is called a *metric* if the following conditions are fulfilled for all $x, y, z \in X$:

- (reflectivity) $d(x, x) = 0$;
- (positivity) $d(x, y) > 0$ for $x \neq y$;
- (symmetry) $d(x, y) = d(y, x)$;
- (triangle inequality) $d(x, y) \leq d(x, z) + d(z, y)$.

A *metric space* (X, d) is a set X equipped with a metric $d\colon X \times X \to \mathbb{R}$.

Definition 3.1.3. The *Euclidean distance* between two points $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ in $\mathbb{R}^n$ is given by the formula,

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

Definition 3.1.4. A dissimilarity matrix $\mathbf{D}$ is called *Euclidean* if there exists a configuration of points in some Euclidean space whose interpoint distances are given by $\mathbf{D}$; that is, if for some m, there exists points $x_1, \ldots, x_n \in \mathbb{R}^m$ such that

$$d_{rs}^2 = (x_r - x_s)^\top (x_r - x_s),$$

where $\top$ denotes the transpose of a matrix.

Definition 3.1.5. A map $\phi : X \to Y$ between metric spaces (X, d_X) and (Y, d_Y) is an *isometric embedding* if $d_Y(\phi(x), \phi(x')) = d_X(x, x')$ for all $x, x' \in X$. The map ϕ is an *isometry* if it is a surjective isometric embedding.

Proximities in the form of similarity matrices also commonly appear in applications of MDS. A reasonable measure of *similarity* $s(a, b)$ must satisfy the following properties [3]:

- (symmetry) $s(a, b) = s(b, a)$,

- (positivity) $s(a, b) > 0$,
- $s(a, b)$ increases as the similarity between a and b increases.

Definition 3.1.6. An $(n \times n)$ matrix $\mathbf{C}$ is called a *similarity matrix* if it is symmetric and

$$c_{rs} \leq c_{rr} \quad \text{for all } r, s.$$

To use the techniques discussed in this thesis, it is necessary to transform the similarities to dissimilarities. The standard transformation from a similarity matrix $\mathbf{C}$ to a dissimilarity matrix $\mathbf{D}$ is defined by

$$d_{rs} = \sqrt{c_{rr} - 2c_{rs} + c_{ss}}.$$

3.2 Inner Product and Normed Spaces

Definition 3.2.1. An *inner product* on a linear space X over the field F ($\mathbb{R}$ or $\mathbb{C}$) is a function $\langle \cdot, \cdot \rangle \colon X \times X \to F$ with the following properties:

- (conjugate symmetry) $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for all $x, y \in X$, where the overline denotes the complex conjugate;
- (linearity in the first argument)

$$\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$$
$$\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$$

  for all $x, y, z \in X$ and $\alpha \in F$;
- (positive-definiteness)

$$\langle x, x\rangle \geq 0$$

$$\langle x, x\rangle = 0 \Leftrightarrow x = \mathbf{0}$$

for all $x \in X$.

An *inner product space* is a vector space equipped with an inner product.

Definition 3.2.2. A *norm* on a linear space X is a function $\|\cdot\|\colon X \to \mathbb{R}$ with the following properties:

- (nonnegative) $\|x\| \geq 0$, for all $x \in X$;
- (homogeneous) $\|\lambda x\| = |\lambda|\|x\|$, for all $x \in X$ and $\lambda \in \mathbb{R}$ (or $\mathbb{C}$);
- (triangle inequality) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$;
- (strictly positive) $\|x\| = 0$ implies $x = 0$.

A *normed linear space* $(X, \|\cdot\|)$ is a linear space X equipped with a norm $\|\cdot\|$.

Definition 3.2.3. A *Hilbert Space* is a complete inner product space with norm defined by the inner product,

$$\|f\| = \sqrt{\langle f, f\rangle}.$$

Definition 3.2.4. The space ℓ^2 is a subspace of $\mathbb{R}^{\mathbb{N}}$ consisting of all sequences $(x_n)_{n=1}^{\infty}$ such that

$$\sum_n |x_n|^2 < \infty.$$

Define an inner product on ℓ^2 for all sequences $(x_n), (y_n) \in \ell^2$ by

$$\langle (x_n), (y_n)\rangle = \sum_{n=1}^{\infty} x_n y_n.$$

The real-valued function $\|\cdot\|_2 \colon \ell^2 \to \mathbb{R}$ defined by $\|(x_n)\|_2 = \left(\sum_{n=1}^{\infty} |x_n|^2\right)^{1/2}$ defines a norm on ℓ^2.

3.3 Metric Measure Spaces

The following introduction to metric measure spaces and the metrics on them is based on [21]. The reader is referred to [21, 22] for detailed descriptions and interpretations of the following concepts in the context of object matching.

Definition 3.3.1. The support of a measure μ on a metric space (Z, d), denoted by $\mathrm{supp}[\mu]$, is the minimal closed subset $Z_0 \subseteq Z$ such that $\mu(Z \backslash Z_0) = 0$.

Given a metric space (X, d_X), by a measure on X we mean a measure on $(X, \mathcal{B}(X))$, where $\mathcal{B}(X)$ is the Borel σ-algebra of X. Given measurable spaces $(X, \mathcal{B}(X))$ and $(Y, \mathcal{B}(Y))$ with measures μ_X and μ_Y, respectively, let $\mathcal{B}(X \times Y)$ be the σ–algebra on $X \times Y$ generated by subsets of the form $A \times B$ with $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Y)$. The product measure $\mu_X \otimes \mu_Y$ is defined to be the unique measure on $(X \times Y, \mathcal{B}(X \times Y))$ such that $\mu_X \otimes \mu_Y(A \times B) = \mu_X(A)\mu_Y(B)$ for all $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Y)$. Furthermore, for $x \in X$, let δ_x^X denote the Dirac measure on X.

Definition 3.3.2. For a measurable map $f : X \to Y$ between two compact metric spaces X and Y, and for μ a measure on X, the *push-forward measure* $f_{\#}\mu$ on Y is given by $f_{\#}\mu(A) = \mu(f^{-1}(A))$ for $A \in \mathcal{B}(Y)$.

Definition 3.3.3. A *metric measure space* is a triple (X, d_X, μ_X) where

- (X, d_X) is a compact metric space, and
- μ_X is a Borel probability measure on X, i.e. $\mu_X(X) = 1$.

In the definition of a metric measure space, it is sometimes assumed that μ_X has full support, namely that $\mathrm{supp}[\mu_X] = X$. When this is done, it is often for notational convenience. For the

metric measure spaces that appear in Section 3.4 we will assume that $\mathrm{supp}[\mu_X] = X$, but we will not make this assumption elsewhere in the document.

Denote by $\mathcal{G}_w$ the collection of all metric measure spaces. Two metric measure spaces (X, d_X, μ_X) and (Y, d_Y, μ_Y) are called *isomorphic* if and only if there exists an isometry $\psi\colon X \to Y$ such that $(\psi)_{\#}\mu_X = \mu_Y$.

3.4 Metrics on Metric Measure Spaces

We describe a variety of notions of distance between two metric spaces or between two metric measure spaces. The content in this section will only be used in Section 8.5. We begin by first defining the Hausdorff distance between two aligned metric spaces, and the Gromov–Hausdorff distance between two unaligned metric spaces. The Gromov–Wasserstein distance is an extension of the Gromov–Hausdorff distance to metric measure spaces. In this section alone, we assume that the metric measure spaces (X, d_X, μ_X) that appear satisfy the additional property that $\mathrm{supp}[\mu_X] = X$.

Definition 3.4.1. Let (Z, d) be a compact metric space. The *Hausdorff distance* between any two closed sets A, $B \subseteq Z$ is defined as

$$d_{\mathcal{H}}^{Z}(A, B) = \max\{ \sup_{a\in A} \inf_{b\in B} d(a, b), \sup_{b\in B} \inf_{a\in A} d(a, b) \}.$$

Definition 3.4.2. The *Gromov–Hausdorff distance* between compact metric spaces X and Y is defined as

$$d_{\mathcal{GH}}(X, Y) = \inf_{Z,f,g} d_{\mathcal{H}}^{Z}(f(X), g(Y)) \tag{3.1}$$

where $f : X \to Z$ and $g : Y \to Z$ are isometric embeddings into the metric space (Z, d).

It is extremely difficult to compute Gromov–Hausdorff distances; indeed the set of all metric spaces Z admitting isometric embeddings from both X and Y is an unbelievably large collection

over which to take an infimum. It is possible to give an equivalent definition of Gromov–Hausdorff distances instead using correspondences (which feel more computable though in practice are still hard to compute.)

Definition 3.4.3. A subset $R \subseteq X \times Y$ is said to be a *correspondence* between sets X and Y whenever $\pi_1(R) = X$ and $\pi_2(R) = Y$, where $\pi_1 \colon X \times Y \to X$ and $\pi_2 \colon X \times Y \to Y$ are the canonical projections.

Let $\mathcal{R}(X, Y)$ denote the set of all possible correspondences between sets X and Y. For metric spaces (X, d_X) and (Y, d_Y), let the *distortion* $\Gamma_{X,Y} \colon X \times Y \times X \times Y \to \mathbb{R}^+$ be given by

$$\Gamma_{X,Y}(x, y, x', y') = |d_X(x, x') - d_Y(y, y')|.$$

Definition 3.4.4. (Alternative form of $d_{\mathcal{GH}}$) One can equivalently (in the sense of equality) define the *Gromov–Hausdorff distance* between compact metric spaces (X, d_X) and (Y, d_Y) as

$$d_{\mathcal{GH}}(X, Y) = \frac{1}{2} \inf_{R \in \mathcal{R}(X,Y)} \sup_{\substack{(x,y) \in R \\ (x',y') \in R}} \Gamma_{X,Y}(x, y, x', y'), \tag{3.2}$$

where R ranges over $\mathcal{R}(X, Y)$.

We now build up the machinery to describe a notion of distance not between metric spaces, but instead between metric measure spaces. Let $\mathcal{C}(Z)$ denote the collection of all compact subsets of Z. We denote the collection of all *weighted objects* in the metric space (Z, d) by

$$\mathcal{C}_w(Z) := \{(A, \mu_A) \mid A \in \mathcal{C}(Z)\},$$

where for each $A \in \mathcal{C}(Z)$, μ_A is a Borel probability measure with $\operatorname{supp}[\mu_A] = A$. Informally speaking, an object in $\mathcal{C}_w(Z)$ is specified not only by the set of points that constitute it, but also

by a distribution of importance over these points. These probability measures can be thought of as acting as weights for each point in the metric space [21].

This following relaxed notion of correspondence between objects is called a matching measure, or a coupling.

Definition 3.4.5. (Matching measure) Let $(A, \mu_A), (B, \mu_B) \in C_w(Z)$. A measure μ on the product space $A \times B$ is a *matching measure* or *coupling* of μ_A and μ_B if

$$\mu(A_0 \times B) = \mu_A(A_0) \quad \text{and} \quad \mu(A \times B_0) = \mu_B(B_0)$$

for all Borel sets $A_0 \subseteq A$ and $B_0 \subseteq B$. Denote by $\mathcal{M}(\mu_A, \mu_B)$ the set of all couplings of μ_A and μ_B.

Proposition 3.4.6. *[21, Lemma 2.2] Let μ_A and μ_B be Borel probability measures on (Z, d), a compact space, with $\operatorname{supp}(\mu_A) = \operatorname{supp}(\mu_B) = Z$. If $\mu \in \mathcal{M}(\mu_A, \mu_B)$, then $\mathcal{R}(\mu) := \operatorname{supp}[\mu]$ belongs to $\mathcal{R}(\operatorname{supp}[\mu_A], \operatorname{supp}[\mu_B])$.*

Definition 3.4.7 (Wasserstein–Kantorovich–Rubinstein distances between measures)**.** For each $p \geq 1$, the following family of distances on $\mathcal{C}_w(Z)$ known as the *Wasserstein distances*, where (Z, d) is a compact metric space:

$$d_{W,p}^Z(A, B) = \left(\inf_{\mu \in \mathcal{M}(\mu_A, \mu_B)} \int_{A \times B} d(a, b)^p \, \mathrm{d}\mu(a, b) \right)^{1/p},$$

for $1 \leq p < \infty$, and

$$d_{W,\infty}^Z(A, B) = \inf_{\mu \in \mathcal{M}(\mu_A, \mu_B)} \sup_{(a,b) \in \mathcal{R}(\mu)} d(a, b).$$

In [21], Mémoli introduced and studied a metric $\mathfrak{D}_p$ on $\mathcal{G}_w$, defined below. An alternative metric $\mathcal{G}_p$ on $\mathcal{G}_w$ is defined and studied by Strum in [34]. The two distances are not equal. Theorem 5.1

of [21] proves that $\mathcal{G}_p \geq \mathfrak{D}_p$ for $1 \leq p \leq \infty$ and that $\mathcal{G}_\infty = \mathfrak{D}_\infty$, where for $p < \infty$ the equality does not hold in general.

For $p \in [1, \infty)$ and $\mu \in \mathcal{M}(\mu_X, \mu_Y)$, let

$$\mathbf{J}_p(\mu) = \frac{1}{2}\left(\int_{X\times Y}\int_{X\times Y} (\Gamma_{X,Y}(x,y,x',y'))^p \mu(\mathrm{d}x \times dy)\mu(\mathrm{d}x' \times \mathrm{d}y')\right)^{\frac{1}{p}},$$

and also let

$$\mathbf{J}_\infty(\mu) = \frac{1}{2} \sup_{\substack{x,x'\in X\\ y,y'\in Y\\ (x,y),(x',y')\in\mathcal{R}(\mu)}} \Gamma_{X,Y}(x,y,x',y').$$

Definition 3.4.8. For $1 \leq p \leq \infty$, define the *Gromov–Wasserstein distance* $\mathfrak{D}_p$ between two metric measure spaces X and Y as

$$\mathfrak{D}_p(X,Y) = \inf_{\mu\in\mathcal{M}(\mu_X,\mu_Y)} \mathbf{J}_p(\mu).$$

In Section 8.5, we pose some questions relating the Gromov–Wasserstein distance to notions of convergence of MDS for possibly infinite metric measure spaces.

Chapter 4

The Theory of Multidimensional Scaling

Multidimensional scaling (MDS) is a set of statistical techniques concerned with the problem of constructing a configuration of n points in a Euclidean space using information about the dissimilarities between the n objects. The dissimilarities between objects need not be based on Euclidean distances; they can represent many types of dissimilarities. The goal of MDS is to map the objects $x_1, \ldots, x_n$ to a configuration (or embedding) of points $f(x_1), \ldots, f(x_n)$ in $\mathbb{R}^m$ in such a way that the given dissimilarities $d(x_i, x_j)$ are well-approximated by the Euclidean distance $\|f(x_i) - f(x_j)\|_2$. The different notions of approximation give rise to the different types of MDS.

If the dissimilarity matrix can be realized exactly as the distance matrix of some set of points in $\mathbb{R}^m$ (i.e. if the dissimilarity matrix is *Euclidean*), then MDS will find such a realization. Furthermore, MDS can be used to identify the minimum such Euclidean dimension m admitting an isometric embedding. However, some dissimilarity matrices or metric spaces are inherently non-Euclidean (cannot be embedded into $\mathbb{R}^m$ for any m). When a dissimilarity matrix is not Euclidean, then MDS produces a mapping into $\mathbb{R}^m$ that distorts the interpoint pairwise distances as little as possible. Though we introduce MDS below, the reader is also referred to [3, 10, 14] for more complete introductions to MDS.

4.1 Types of Multidimensional Scaling

There are several types of MDS, and they differ mostly in the loss function they minimize. In general, there are two dichotomies (the following discussion is from [9]) :

1. Kruskal–Shepard distance scaling versus classical Torgerson–Gower inner-product scaling: In distance scaling, dissimilarities are fitted by distances, whereas classical scaling transforms the dissimilarities to a form that is naturally fitted by inner products.

2. Metric scaling versus nonmetric scaling: Metric scaling uses the actual values of the dissimilarities, while nonmetric scaling effectively uses only their ranks, i.e., their orderings [17,28, 29]. Nonmetric MDS is realized by estimating an optimal monotone transformation $f(d_{ij})$ of the dissimilarities while simultaneously estimating the configuration.

There are two main differences between classical and distance scaling. First, inner products rely on an origin, while distances do not. So, a set of inner products determines uniquely a set of distances, but a set of distances determines a set of inner products only modulo change of origin. To avoid arbitrariness, one constrains classical scaling to mean-centered configurations. Second, distance scaling requires iterative minimization while classical scaling can be solved in a single step by computing an inexpensive eigendecomposition. The different types of MDS arise from different combinations of {metric, nonmetric} with {distance, classical}. The reader is referred to [9] for further discussion on the topic.

Two common loss functions are known as the Strain and Stress functions. We call "Strain" any loss function that measures the lack of fit between inner products $\langle f(x_i), f(x_j)\rangle$ of the configuration points in $\mathbb{R}^m$ and the inner-product data b_{ij} of the given data points. The following is an example of a Strain function, where f is the MDS embedding map:

$$\mathrm{Strain}(f) = \sum_{i,j} \left(b_{ij} - \langle f(x_i), f(x_j)\rangle\right)^2.$$

We call "Stress" any loss function that measures the lack of fit between the Euclidean distances $\hat{d}_{ij}$ of the configuration points and the given proximities δ_{ij}. The general form of Stress [10, 36] is

$$\mathrm{Stress}(f) = \sqrt{\frac{\sum_{i,j} \left(h(\delta_{ij}) - \hat{d}_{ij}\right)^2}{scale}},$$

where f is the MDS embedding map, where h is a smoothing function of the data, and where the '$scale$' component refers to a constant scaling factor, used to keep the value of Stress in the

convenient range between 0 and 1. The choice of h depends on the type of MDS needed. In metric scaling, h is the identity map, which means that the raw input proximity data is compared directly to the mapped distances. In non-metric scaling, however, h is usually an arbitrary monotone function that can be optimized over. The reader is referred to [8,10,36] for descriptions of the most common forms of Stress.

4.2 Metric Multidimensional Scaling

Suppose we are given n objects $x_1, \ldots, x_n$ with the dissimilarities d_{ij} between them, for $i, j = 1, \ldots, n$. Metric MDS attempts to find a set of points $f(x_1), \ldots, f(x_n)$ in a Euclidean space of some dimension where each point represents one of the objects, and the distances between points $\hat{d}_{ij}$ are such that

$$\hat{d}_{ij} \approx h(d_{ij}).$$

Here h is typically the identity function, but could also be a continuous parametric monotone function that attempts to transform the dissimilarities to a distance-like form [10]. Assuming that all proximities are already in a satisfactory distance-like form, the aim is to find a mapping f, for which d_{ij} is approximately equal to $\hat{d}_{ij}$, for all i, j. The two main metric MDS methods are classical scaling and least squares scaling. We will introduce both, with most emphasis placed on the former.

4.2.1 Classical Scaling

Classical multidimensional scaling (cMDS) is also known as Principal Coordinates Analysis (PCoA), Torgerson Scaling, or Torgerson–Gower scaling. The cMDS algorithm minimizes a Strain function, and one of the main advantages of cMDS is that its algorithm is algebraic and not iterative. Therefore, it is simple to implement, and is guaranteed to discover the optimal configura-

tion in $\mathbb{R}^m$. In this section, we describe the algorithm of cMDS, and then we discuss its optimality properties and goodness of fit.

Let $\mathbf{D} = (d_{ij})$ be an $n \times n$ dissimilarity matrix. Let $\mathbf{A} = (a_{ij})$, where $a_{ij} = -\frac{1}{2}d_{ij}^2$. Define

$$\mathbf{B} = \mathbf{HAH}, \tag{4.1}$$

where $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top$ is the centering matrix of size n ($\mathbf{1}$ is a column-vector of n ones and $\top$ denotes matrix transpose). Multiplying $\mathbf{A}$ by the matrix $\mathbf{H}$ on either side has the effect of double-centering the matrix. Indeed, we have

$$b_{rs} = a_{rs} - \overline{a_{r\cdot}} - \overline{a_{\cdot s}} + \overline{a_{\cdot\cdot}},$$

where $\overline{a_{r\cdot}} = \frac{1}{n}\sum_{s=1}^{n} a_{rs}$ is the average of row r, where $\overline{a_{\cdot s}} = \frac{1}{n}\sum_{r=1}^{n} a_{rs}$ is the average of column s, and where $\overline{a_{\cdot\cdot}} = \frac{1}{n^2}\sum_{r,s=1}^{n} a_{rs}$ is the average entry in the matrix. Since $\mathbf{D}$ is a symmetric matrix, it follows that $\mathbf{A}$ and $\mathbf{B}$ are each symmetric, and therefore $\mathbf{B}$ has n real eigenvalues.

Assume for convenience that there are at least m positive eigenvalues for matrix $\mathbf{B}$, where $m \leq n$. By the spectral theorem of symmetric matrices, let $\mathbf{B} = \mathbf{\Gamma\Lambda\Gamma}^\top$ with $\mathbf{\Gamma}$ containing unit-length eigenvectors of $\mathbf{B}$ as its columns, and with the diagonal matrix $\mathbf{\Lambda}$ containing the eigenvalues of $\mathbf{B}$ in decreasing order along its diagonal. Let $\mathbf{\Lambda}_m$ be the $m \times m$ diagonal matrix of the largest m eigenvalues sorted in descending order, and let $\mathbf{\Gamma}_m$ be the $n \times m$ matrix of the corresponding m eigenvectors in $\mathbf{\Gamma}$. Then the coordinates of the MDS embedding into $\mathbb{R}^m$ are given by the $n \times m$ matrix $\mathbf{X} = \mathbf{\Gamma}_m\mathbf{\Lambda}_m^{1/2}$. More precisely, the MDS embedding consists of the n points in $\mathbb{R}^m$ given by the n rows of $\mathbf{X}$.

The procedure for classical MDS can be summarized in the following steps.

1. Compute the matrix $\mathbf{A} = (a_{ij})$, where $a_{ij} = -\frac{1}{2}d_{ij}^2$.

2. Apply double-centering to $\mathbf{A}$: Define $\mathbf{B} = \mathbf{HAH}$, where $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top$.

3. Compute the eigendecomposition of $\mathbf{B} = \mathbf{\Gamma \Lambda \Gamma}^\top$.

4. Let $\mathbf{\Lambda}_m$ be the matrix of the largest m eigenvalues sorted in descending order, and let $\mathbf{\Gamma}_m$ be the matrix of the corresponding m eigenvectors. Then, the coordinate matrix of classical MDS is given by $\mathbf{X} = \mathbf{\Gamma}_m \mathbf{\Lambda}_m^{1/2}$.

The following is a fundamental criterion in determining whether the dissimilarity matrix $\mathbf{D}$ is Euclidean or not.

Theorem 4.2.1. *[3, Theorem 14.2.1] Let $\mathbf{D}$ be a dissimilarity matrix, and define $\mathbf{B}$ by equation (4.1). Then $\mathbf{D}$ is Euclidean if and only if $\mathbf{B}$ is a positive semi-definite matrix.*

In particular, if $\mathbf{B}$ is positive semi-definite of rank m, then a perfect realization of the dissimilarities can be found by a collection of points in m-dimensional Euclidean space.

If we are given a Euclidean matrix $\mathbf{D}$, then the classical solution to the MDS problem in k dimensions has the following optimal property:

Theorem 4.2.2. *[3, Theorem 14.4.1] Let $\mathbf{D}$ be a Euclidean distance matrix corresponding to a configuration $\mathbf{X}$ in $\mathbb{R}^m$, and fix k $(1 \leq k \leq m)$. Then amongst all projections $\mathbf{XL_1}$ of $\mathbf{X}$ onto k-dimensional subspaces of $\mathbb{R}^m$, the quantity $\sum_{r,s=1}^{n} (d_{rs}^2 - \hat{d}_{rs}^2)$ is minimized when $\mathbf{X}$ is projected onto its principal coordinates in k dimensions.*

This theorem states that in this setting, MDS minimizes the sum of squared errors in distances, over all possible projections. In the following paragraph, we show that an analogous result is true for the sum of squared errors in inner-products, i.e., for the loss function Strain.

Let $\mathbf{D}$ be a dissimilarity matrix, and let $\mathbf{B} = \mathbf{HAH}$. A measure of the goodness of fit of MDS, even in the case when $\mathbf{D}$ is not Euclidean, can be obtained as follows [3]. If $\hat{\mathbf{X}}$ is a fitted configuration in $\mathbb{R}^m$ with centered inner-product matrix $\hat{\mathbf{B}}$, then a measure of the discrepancy

between $\mathbf{B}$ and $\hat{\mathbf{B}}$ is the following Strain function [20],

$$\mathrm{tr}((\mathbf{B}-\hat{\mathbf{B}})^2) = \sum_{i,j=1}^{n} (b_{i,j} - \hat{b}_{i,j})^2. \tag{4.2}$$

Theorem 4.2.3. *[3, Theorem 14.4.2] Let* $\mathbf{D}$ *be a dissimilarity matrix (not necessarily Euclidean). Then for fixed* m*, the* Strain *function in* (4.2) *is minimized over all configurations* $\hat{\mathbf{X}}$ *in* m *dimensions when* $\hat{\mathbf{X}}$ *is the classical solution to the MDS problem.*

The reader is referred to [10, Section 2.4] for a summary of a related optimization with a different normalization, due to Sammon [25].

4.2.2 Distance Scaling

Some popular metric MDS methods that attempt to minimize distances (versus inner-products) include

- least squares scaling, which minimizes a variation of loss functions, and
- metric Scaling by Majorising a Complicated Function (SMACOF), which minimizes a form of the Stress function.

The reader is referred to [10] for a description of least squares scaling and to [8, 10] for the theory of SMACOF and The Majorisation Algorithm.

4.3 Non-Metric Multidimensional Scaling

Non-Metric Multidimensional Scaling assumes that only the ranks or orderings of the dissimilarites are known. Hence, this method produces a map which tries to reproduce these ranks and not the observed or actual dissimilarities. Thus, only the ordering of the dissimilarities is relevant to the methods of approximations. In [10, Chapter 3], the authors present the underlying theory of non-metric multidimensional scaling developed in the 1960s, including Kruskal's method. Other

methods that fall under Non-Metric Multidimensional Scaling include Non-Metric SMACOF and Sammon Mapping.

4.4 Simple Examples: Visualization

In this section, we consider three simple dissimilarity matrices (input of MDS) and their Euclidean embeddings in $\mathbb{R}^2$ or $\mathbb{R}^3$ (output of MDS). The first two are Euclidean distance matrices, whereas the third is non-Euclidean.

1. Consider the following dissimilarity matrix,

$$D_1 = \begin{pmatrix} 0 & 6 & 8 \\ 6 & 0 & 10 \\ 8 & 10 & 0 \end{pmatrix},$$

which is an example of a dissimilarity matrix that can be isometrically embedded in $\mathbb{R}^2$ (Figure 4.1a) but not in $\mathbb{R}^1$.

2. Consider the following dissimilarity matrix,

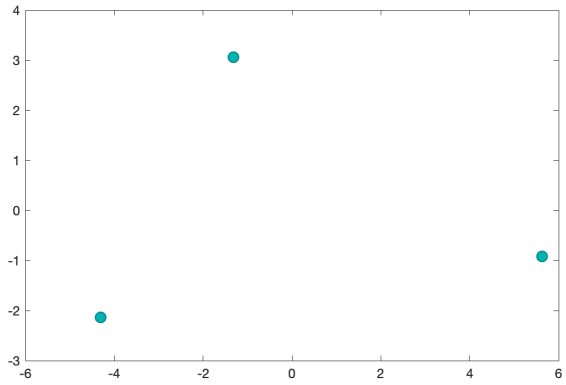
$$D_2 = \begin{pmatrix} 0 & 1 & 1 & \sqrt{2} & 1 \\ 1 & 0 & \sqrt{2} & 1 & 1 \\ 1 & \sqrt{2} & 0 & 1 & 1 \\ \sqrt{2} & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix},$$

which is an example of a dissimilarity matrix that can be isometrically embedded in $\mathbb{R}^3$ (Figure 4.1b) but not in $\mathbb{R}^2$.

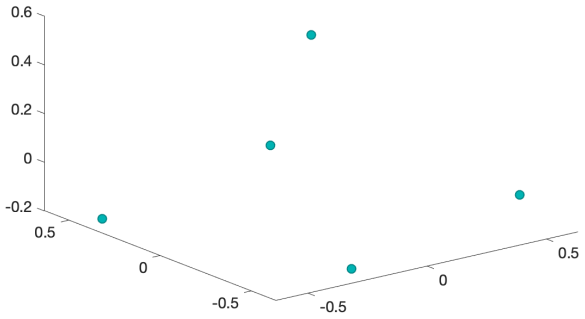
3. Consider the following dissimilarity matrix,

$$D_3 = \begin{pmatrix} 0 & 2 & 2 & 1 \\ 2 & 0 & 2 & 1 \\ 2 & 2 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix},$$

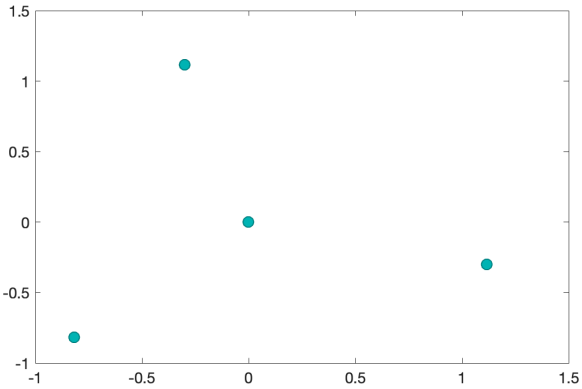
which is an example of a dissimilarity matrix that cannot be isometrically embedded into any Euclidean space. Indeed, label the points in the metric space x_1, x_2, x_3, x_4 in order of their row/column in D_3. In any isometric embedding in $\mathbb{R}^n$, the points x_1, x_2, x_3 would need to get mapped to an equilateral triangle. Then the point x_4 would need to get mapped to the midpoint of each edge in this triangle, which is impossible in Euclidean space. Figure 4.1c shows the embedding of this metric space in $\mathbb{R}^2$.



(a) MDS embedding of D_1 into $\mathbb{R}^2$.



(b) MDS embedding of D_2 into $\mathbb{R}^3$.



(c) MDS embedding of D_3 into $\mathbb{R}^2$.

Figure 4.1: MDS embeddings into $\mathbb{R}^2$ or $\mathbb{R}^3$ of the three dissimilarity matrices D_1, D_2, and D_3.

Chapter 5

Operator Theory

This section serves to inform the reader on some of the concepts in infinite-dimensional linear algebra and operator theory used throughout our work.

5.1 Kernels and Operators

We denote by $L^2(X, \mu)$ the set of square integrable L^2-functions with respect to the measure μ. We note that $L^2(X, \mu)$ is furthermore a Hilbert space, after equipping it with the inner product given by

$$\langle f, g \rangle = \int_X f g \, d\mu.$$

Definition 5.1.1. A measurable function f on $X \times X$ is said to be *square-integrable* if satisfies the following three conditions [15, 33]:

- $f(x, s)$ is a measurable function of $(x, s) \in X \times X$, with

$$\int_X \int_X |f(x, s)|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}s) < \infty;$$

- for each $s \in X$, the function $f(x, s)$ is a measurable function in x, with

$$\int_X |f(x, s)|^2 \mu(\mathrm{d}x) < \infty;$$

- for each $x \in X$, the function $f(x, s)$ is a measurable function in s, with

$$\int_X |f(x, s)|^2 \mu(\mathrm{d}s) < \infty.$$

The L^2-norm of a square-integrable function is given by

$$\int_X \int_X |f(x,s)|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}s) < \infty.$$

We denote by $L^2_{\mu\otimes\mu}(X \times X)$ the set of square integrable functions with respect to the measure $\mu \otimes \mu$.

Definition 5.1.2. A set $\{f_i\}_{i\in\mathbb{N}}$ of real-valued functions $f_i \in L^2(X,\mu)$ is said to be *orthonormal* if

$$\langle f_i, f_j \rangle = \delta_{ij}.$$

When it is clear from the context, we will simply write $L^2(X)$ and $L^2(X \times X)$ instead of $L^2(X,\mu)$ and $L^2_{\mu\otimes\mu}(X \times X)$, respectively.

In this context, a *real-valued L^2-kernel* $K\colon X \times X \to \mathbb{R}$ is a continuous measurable square-integrable function i.e. $K \in L^2_{\mu\otimes\mu}(X \times X)$. Most of the kernels that we define in our work are symmetric.

Definition 5.1.3. A kernel K is *symmetric* (or *complex symmetric* or *Hermitian*) if

$$K(x,s) = \overline{K(s,x)} \quad \text{for all } x, s \in X,$$

where the overline denotes the complex conjuguate. In the case of a real kernel, the symmetry reduces to the equality

$$K(x,s) = K(s,x).$$

Definition 5.1.4. A symmetric function $K : X \times X \to \mathbb{R}$ is called a *positive semi-definite kernel* on X if

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i,x_j) \geq 0$$

holds for any $m \in \mathbb{N}$, any $x_1, \ldots, x_m \in X$, and any $c_1, \ldots, c_m \in \mathbb{R}$.

We remark that the above definition is equivalent to saying that for all vectors $c \in \mathbb{R}^m$, we have

$$c^T K c \geq 0, \quad \text{where } K_{ij} = K(x_i, x_j).$$

At least in the case when X is a compact subspace of $\mathbb{R}^m$ (and probably more generally), we have the following equivalent definition.

Definition 5.1.5. Let X be a compact subspace of $\mathbb{R}^m$, and let $K \in L^2_{\mu\otimes\mu}(X \times X)$ be a real-valued symmetric kernel. Then $K(x, s)$ is a *positive semi-definite kernel* on X if

$$\int_{X\times X} K(x, s) f(x) f(s) d(\mu \otimes \mu)(x, s) \geq 0$$

for any $f \in L^2(X, \mu)$.

Definition 5.1.6. Let $\mathbb{C}^{m\times n}$ denote the space of all $m \times n$ matrices with complex entries. We define an inner product on $\mathbb{C}^{m\times n}$ by

$$(A, B) = \mathrm{Tr}(A^* B),$$

where Tr denotes the trace and * denotes the Hermitian conjugate of a matrix. If $A = (a_{ij})$ and $B = (b_{ij})$, then

$$(A, B) = \sum_{i=1}^{m} \sum_{j=1}^{n} \overline{a_{ij}} b_{ij}.$$

The corresponding norm,

$$\|A\| = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2},$$

is called the *Hilbert–Schmidt norm*.

For $\mathcal{H}_1$ and $\mathcal{H}_2$ Hilbert spaces, we let $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ denote the set of bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. Similarly, for $\mathcal{H}$ a Hilbert space, we let $\mathcal{B}(\mathcal{H})$ denote the set of bounded linear operators from $\mathcal{H}$ to itself.

Definition 5.1.7. Let $T \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. Then there exists a unique operator from $\mathcal{H}_2$ into $\mathcal{H}_1$, denoted by T^* (or T'), such that

$$\langle Tx, y\rangle = \langle x, T^*y\rangle \quad \forall x \in \mathcal{H}_1, y \in \mathcal{H}_2.$$

This operator T^* is called the *adjoint* of T.

Definition 5.1.8. Suppose $T \in \mathcal{B}(\mathcal{H})$. If $T = T^*$, then T is called *self-adjoint*.

Definition 5.1.9. A bounded linear operator $A \in \mathcal{B}(\mathcal{H})$ over a separable Hilbert space $\mathcal{H}$ is said to be in the *trace class* if for some (and hence all) orthonormal bases $\{e_k\}_{k\in\mathbb{N}}$ of $\mathcal{H}$, the sum of positive terms

$$\|A\|_1 = \mathrm{Tr}|A| = \sum_k \langle (A^*A)^{1/2}\, e_k, e_k\rangle$$

is finite. In this case, the trace of A, which is given by the sum

$$\mathrm{Tr}(A) = \sum_k \langle Ae_k, e_k\rangle,$$

is absolutely convergent and is independent of the choice of the orthonormal basis.

Definition 5.1.10. A *Hilbert–Schmidt operator* is a bounded linear operator A on a Hilbert space $\mathcal{H}$ with finite Hilbert–Schmidt norm

$$\|A\|_{HS}^2 = \mathrm{Tr}(A^*A) = \sum_{i\in I} \|Ae_i\|^2,$$

where $\|\cdot\|$ is the norm of $\mathcal{H}$, $\{e_i : i \in I\}$ is an orthonormal basis of $\mathcal{H}$, and Tr is the trace of a nonnegative self-adjoint operator.

Note that the index set need not be countable. This definition is independent of the choice of the basis, and therefore

$$\|A\|_{HS}^2 = \sum_{i,j} |A_{i,j}|^2,$$

where $A_{i,j} = \langle e_i, Ae_j \rangle$. In Euclidean space, the Hilbert–Schmidt norm is also called the Frobenius norm.

When $\mathcal{H}$ is finite-dimensional, every operator is trace class, and this definition of the trace of A coincides with the definition of the trace of a matrix.

Proposition 5.1.11. *If A is bounded and B is trace class (or if A and B are Hilbert–Schmidt), then AB and BA are also trace class, and* $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$.

Definition 5.1.12 (Hilbert–Schmidt Integral Operator)**.** Let (X, Ω, μ) be a σ-finite measure space, and let $K \in L^2_{\mu\otimes\mu}(X \times X)$. Then the integral operator

$$[T_K \phi](x) = \int_X K(x, s)\phi(s)\mu(\mathrm{d}s)$$

defines a linear mapping acting from the space $L^2(X, \mu)$ into itself.

Hilbert–Schmidt integral operators are both continuous (and hence bounded) and compact operators.

Suppose $\{e_i\}_{i\in\mathbb{N}}$ is an orthonormal basis of $L^2(X, \mu)$. Define $e_{nm}(x, s) = e_n(s)e_m(x)$. Then, $(e_{mn})_{m,n\in\mathbb{N}}$ forms an orthonormal basis of $L^2_{\mu\otimes\mu}(X \times X)$.

Proposition 5.1.13. $\|T_K\|_{HS}^2 = \|K\|_{L^2(X\times X)}^2$.

Proof. By Definition 5.1.10, we have

$$
\begin{aligned}
\|T_K\|_{HS}^2 &= \sum_n \|T_K e_n\|_{L^2(X)} \\
&= \sum_{n,m} |\langle e_m, T_K e_n \rangle_{L^2(X)}|^2 \\
&= \sum_{n,m} \left| \left\langle e_m, \int K(x,s) e_n(x) \mu(\mathrm{d}x) \right\rangle_{L^2(X)} \right|^2 \\
&= \sum_{n,m} \left| \int_X \int_X K(x,s) e_n(x) e_m(s) \mu(\mathrm{d}x) \mu(\mathrm{d}s) \right|^2 \\
&= \sum_{n,m} |\langle K, e_{nm} \rangle_{L^2(X\times X)}|^2 \\
&= \|K\|_{L^2(X\times X)}^2.
\end{aligned}
$$

□

Definition 5.1.14. A Hilbert–Schmidt integral operator is a *self-adjoint operator* if and only if $K(x,y) = \overline{K(y,x)}$ (i.e K is a symmetric kernel) for almost all $(x,y) \in X \times X$ (with respect to $\mu \times \mu$).

Remark 5.1.15. Suppose T_K a self-adjoint Hilbert-Schmidt integral operator. By Definition 5.1.10, we have

$$
\|T_K\|_{HS}^2 = \mathrm{Tr}(T_K^* T_K) = \mathrm{Tr}((T_K)^2),
$$

and by Proposition 5.1.13

$$
\|T_K\|_{HS}^2 = \|K\|_{L^2(X\times X)}^2 = \int_X \int_X |K(x,s)|^2 \mu(\mathrm{d}x) \mu(\mathrm{d}s).
$$

These properties will be useful in defining a Strain function for MDS of infinite metric measure spaces in Section 6.4.

Definition 5.1.16. A bounded self-adjoint operator A on a Hilbert space $\mathcal{H}$ is called a *positive semi-definite operator* if $\langle Ax, x\rangle \geq 0$ for any $x \in \mathcal{H}$.

It follows that for every positive semi-definite operator A, the inner product $\langle Ax, x\rangle$ is real for every $x \in \mathcal{H}$. Thus, the eigenvalues of A, when they exist, are real.

Definition 5.1.17. A linear operator $T\colon L^2(X,\mu) \to L^2(X,\mu)$ is said to be *orthogonal* (in $\mathbb{R}$) or *unitary* (in $\mathbb{C}$) if $\langle f, g\rangle = \langle T(f), T(g)\rangle$ for all f and g in $L^2(X,\mu)$.

Definition 5.1.18. Suppose $h, g \in L^2(X,\mu)$. Define the linear operator $h \otimes g\colon L^2(X,\mu) \to L^2(X,\mu)$ as follows

$$(h \otimes g)(f) = \langle g, f\rangle h.$$

Then $h\otimes g$ is a Hilbert–Schmidt integral operator associated with the L^2-kernel $K(x,s) = h(x)g(s)$.

Suppose $\{e_n\}_{n\in\mathbb{N}}$ and $\{f_n\}_{n\in\mathbb{N}}$ are two distinct orthonormal bases for $L^2(X,\mu)$. Define the linear operator $G\colon L^2(X,\mu) \to L^2(X,\mu)$ by

$$G(\cdot) = \sum_n f_n \otimes e_n = \sum_n \langle e_n, \cdot\rangle f_n.$$

Proposition 5.1.19. *G is an orthogonal operator.*

Proof. For any $\phi, \psi \in L^2(x,\mu)$, we have

$$\begin{aligned}\langle G(\phi), G(\psi)\rangle &= \left\langle \sum_i \langle e_i, \phi\rangle f_i, \sum_j \langle e_j, \psi\rangle f_j \right\rangle \\ &= \sum_{i,j} \langle \langle e_i, \phi\rangle f_i, \langle e_j, \psi\rangle f_j\rangle \\ &= \sum_{i,j} \langle e_i, \phi\rangle \langle e_j, \psi\rangle \langle f_i, f_j\rangle \\ &= \sum_i \langle e_i, \phi\rangle \langle e_i, \psi\rangle \\ &= \langle \phi, \psi\rangle.\end{aligned}$$

Therefore, G is an orthogonal operator. □

More generally, if $\{e_n\}_{n\in\mathbb{N}}$ and $\{f_n\}_{n\in\mathbb{N}}$ are any bases for $L^2(X,\mu)$, then $G = \sum_n f_n \otimes e_n$ is a change-of-basis operator (from the $\{e_n\}_{n\in\mathbb{N}}$ basis to the $\{f_n\}_{n\in\mathbb{N}}$ basis), satisfying $G(e_i) = f_i$ for all i.

5.2 The Spectral Theorem

Definition 5.2.1. A complex number $\lambda \in \mathbb{C}$ is an *eigenvalue* of $T \in \mathcal{B}(\mathcal{H})$ if there exists a non-zero vector $x \in \mathcal{H}$ such that $Tx = \lambda x$. The vector x is called an *eigenvector* for T corresponding to the eigenvalue λ. Equivalently, λ is an eigenvalue of T if and only if $T - \lambda I$ is not one-to-one.

Theorem 5.2.2 (Spectral theorem on compact self-adjoint operators)**.** *Let $\mathcal{H}$ be a not necessarily separable Hilbert space, and suppose $T \in \mathcal{B}(\mathcal{H})$ is compact self-adjoint operator. Then T has at most a countable number of nonzero eigenvalues $\lambda_n \in \mathbb{R}$, with a corresponding orthonormal set $\{e_n\}$ of eigenvectors such that*

$$T(\cdot) = \sum_n \lambda_n \langle e_n, \cdot\rangle\, e_n.$$

Furthermore, the multiplicity of each nonzero eigenvalue is finite, zero is the only possible accumulation point of $\{\lambda_n\}$, and if the set of non-zero eigenvalues is infinite then zero is necessarily an accumulation point.

A fundamental theorem that characterizes positive semi-definite kernels is the Generalized Mercer's Theorem, which states the following:

Theorem 5.2.3. *[18, Lemma 1] Let X be a compact topological Hausdorff space equipped with a finite Borel measure μ, and let $K\colon X \times X \to \mathbb{C}$ be a continuous positive semi-definite kernel. Then there exists a scalar sequence $\{\lambda_n\} \in \ell_1$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and an orthonormal system $\{\phi_n\}$ in $L^2(X, \mu)$ consisting of continuous functions only, such that the expansion*

$$K(x, s) = \sum_{n=1}^{\infty} \lambda_n \phi_n(x) \phi_n(s), \quad x, s \in \operatorname{supp}(\mu) \tag{5.1}$$

converges uniformly.

Therefore, we have the following. A Hilbert–Schmidt integral operator

$$[T_K \phi](x) = \int_X K(x, s) \phi(s) \mu(\mathrm{d}s) \tag{5.2}$$

associated to a symmetric positive semi-definite L^2-kernel K, is a positive-semi definite operator. Moreover, the eigenvalues of T_K can be arranged in non-increasing order, counting them according to their algebraic multiplicities: $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.

Chapter 6

MDS of Infinite Metric Measure Spaces

Classical multidimensional scaling can be described either as a Strain-minimization problem, or as a linear algebra algorithm involving eigenvalues and eigenvectors. Indeed, one of the main theoretical results for cMDS is that the linear algebra algorithm solves the corresponding Strain-minimization problem (see Theorem 4.2.3). In this section, we describe how to generalize both of these formulations to infinite metric measure spaces. This will allow us to discuss cMDS of the entire circle for example, without needing to restrict attention to finite subsets thereof.

6.1 Proposed Approach

Suppose we are given a bounded (possibly infinite) metric measure space (X, d_X, μ), where d_X is a real-valued L^2-function on $X \times X$ with respect to the measure $\mu \otimes \mu$. By *bounded* we mean that d_X is a bounded L^2-kernel, i.e. there exists a constant $C \in \mathbb{R}$ such that $d(x, s) \leq C$ for all $x, s \in X$. When it is clear from the context, the triple (X, d_X, μ) will be denoted by only X. Even if X is not Euclidean, we hope that the metric d_X can be approximately represented by a Euclidean metric $d_{\hat{X}} : \hat{X} \times \hat{X} \to \mathbb{R}$ on a space $\hat{X} \subseteq \ell^2$ or $\hat{X} \subseteq \mathbb{R}^m$, where perhaps the Euclidean space is of low dimension (often $m = 2$ or 3).

A metric space (X, d_X) is said to be *Euclidean* if (X, d_X) can be isometrically embedded into $(\ell^2, \|\cdot\|_2)$. That is, (X, d_X) is Euclidean if there exists an isometric embedding $f\colon X \to \ell^2$, meaning $\forall x, s \in X$, we have that $d_X(x, s) = d_{\ell^2}(f(x), f(s))$. Furthermore, we call a metric measure space (X, d_X, μ_X) *Euclidean* if its underlying metric space (X, d_X) is. As will be discussed below, X could also be Euclidean in the finite-dimensional sense, meaning that there is an isometric embedding $f\colon X \to \mathbb{R}^m$.

6.2 Relation Between Distances and Inner Products

The following discussion is carried analogously to the arguments presented in [23] (for spaces of finitely many points). Consider the bounded metric measure space (X, d, μ) where $X \subseteq \ell^2$ (or $\mathbb{R}^m$) and $d\colon X \times X \to \mathbb{R}$ the Euclidean distance. We have the following relation between distances and inner-products in ℓ^2 (or $\mathbb{R}^m$),

$$d^2(x,s) = \langle x-s, x-s\rangle = \langle x,x\rangle + \langle s,s\rangle - 2\langle x,s\rangle = \|x\|^2 + \|s\|^2 - 2\langle x,s\rangle.$$

This implies

$$\langle x,s\rangle = -\frac{1}{2}\left(d^2(x,s) - \|x\|^2 - \|s\|^2\right). \tag{6.1}$$

Furthermore, if we let let $\bar{x} = \int_X w\ \mu(\mathrm{d}w)$ denote the center of the space X, then

$$\begin{aligned}
d^2(x,\bar{x}) =& \langle x,x\rangle + \langle \bar{x},\bar{x}\rangle - 2\langle x,\bar{x}\rangle \\
=& \|x\|^2 + \int_X\int_X \langle w,z\rangle\ \mu(\mathrm{d}z)\mu(\mathrm{d}w) - 2\int_X \langle x,z\rangle\ \mu(\mathrm{d}z) \\
=& \|x\|^2 + \frac{1}{2}\int_X\int_X \left(\|w\|^2 + \|z\|^2 - d^2(w,z)\right)\mu(\mathrm{d}z)\mu(\mathrm{d}w) - \int_X \left(\|x\|^2 + \|z\|^2 - d^2(x,z)\right)\mu(\mathrm{d}z) \\
=& \int_X d^2(x,z)\ \mu(\mathrm{d}z) - \frac{1}{2}\int_X\int_X d^2(w,z)\ \mu(\mathrm{d}z)\mu(\mathrm{d}z).
\end{aligned}$$

Without loss of generality, we can assume our space X is centered at the origin (i.e $\bar{x} = \mathbf{0}$). This implies that $\|x\|^2 = d^2(x,\mathbf{0}) = d^2(x,\bar{x})$. Therefore, equation (6.1) can be written as

$$\langle x,s\rangle = -\frac{1}{2}\left(d^2(x,s) - \int_X d^2(x,z)\ \mu(\mathrm{d}z) - \int_X d^2(w,s)\ \mu(\mathrm{d}w) + \int_X\int_X d^2(w,z)\ \mu(\mathrm{d}w)\mu(\mathrm{d}z)\right). \tag{6.2}$$

Now, let (X, d, μ) be any metric measure space where d is an L^2-function on $X \times X$ with respect to the measure $\mu \otimes \mu$. Define $K_A(x, s) = -\frac{1}{2}d^2(x, s)$ and define K_B as

$$K_B(x, s) = K_A(x, s) - \int_X K_A(x, z)\, \mu(\mathrm{d}z) - \int_X K_A(w, s)\, \mu(\mathrm{d}w) + \int_{X\times X} K_A(w, z)\, \mu(\mathrm{d}w \times \mathrm{d}z).$$

Proposition 6.2.1. *Given a metric measure space (X, d, μ), where $d \in L^2_{\mu\otimes\mu}(X \times X)$, construct K_A and K_B as defined above. Then we have the relation*

$$d^2(x, s) = K_B(x, x) + K_B(s, s) - 2K_B(x, s).$$

Proof. We have

$$\begin{aligned}
&K_B(x, x) + K_B(s, s) - 2K_B(x, s) \\
&= K_A(x, x) - \int_X K_A(x, z)\mu(\mathrm{d}z) - \int_X K_A(w, x)\mu(\mathrm{d}w) + \int_{X\times X} K_A(w, z)\mu(\mathrm{d}w)\mu(\mathrm{d}z) \\
&+ K_A(s, s) - \int_X K_A(s, z)\mu(\mathrm{d}z) - \int_X K_A(w, s)\mu(\mathrm{d}w) + \int_{X\times X} K_A(w, z)\mu(\mathrm{d}w)\mu(\mathrm{d}z) \\
&- 2K_A(x, s) + 2\int_X K_A(x, z)\mu(\mathrm{d}z) + 2\int_X K_A(w, s)\mu(\mathrm{d}w) - 2\int_{X\times X} K_A(w, z)\mu(\mathrm{d}w)\mu(\mathrm{d}z) \\
&= -2K_A(x, s) = d^2(x, s).
\end{aligned}$$

Indeed, the intermediate steps follow since $K_A(x, x) = K_A(s, s) = 0$, and since by the symmetry of K_A we have $\int_X K_A(x, z)\mu(\mathrm{d}z) = \int_X K_A(w, x)\mu(\mathrm{d}w)$ and $\int_X K_A(s, z)\mu(\mathrm{d}z) = \int_X K_A(w, s)\mu(\mathrm{d}w)$. □

6.3 MDS on Infinite Metric Measure Spaces

In this section, we explain how multidimensional scaling generalizes to possibly infinite metric measure spaces that are bounded. We remind the reader that by definition, all of the metric measure spaces we consider are equipped with probability measures.

Let (X, d, μ) be a bounded metric measure space, where d is a real-valued L^2-function on $X \times X$ with respect to the measure $\mu \otimes \mu$. We propose the following MDS method on infinite metric measure spaces:

(i) From the metric d, construct the kernel $K_A \colon X \times X \to \mathbb{R}$ defined as $K_A(x, s) = -\frac{1}{2}d^2(x, s)$.

(ii) Obtain the kernel $K_B \colon X \times X \to \mathbb{R}$ defined as

$$K_B(x,s) = K_A(x,s) - \int_X K_A(w,s)\mu(\mathrm{d}w) - \int_X K_A(x,z)\mu(\mathrm{d}z) + \int_{X\times X} K_A(w,z)\mu(\mathrm{d}w \times \mathrm{d}z). \tag{6.3}$$

Assume $K_B \in L^2(X \times X)$. Define $T_{K_B} \colon L^2(X) \to L^2(X)$ as

$$[T_{K_B}\phi](x) = \int_X K_B(x,s)\phi(s)\mu(\mathrm{d}s).$$

Note that, kernels K_A and K_B are symmetric, since d is a metric.

(iii) Let $\lambda_1 \geq \lambda_2 \geq \ldots$ denote the eigenvalues of T_{K_B} with corresponding eigenfunctions $\phi_1, \phi_2, \ldots$, where the $\phi_i \in L^2(X)$ are real-valued functions. Indeed, $\{\phi_i\}_{i\in\mathbb{N}}$ forms an orthonormal system of $L^2(X)$.

(iv) Define $K_{\hat{B}}(x,s) = \sum_{i=1}^{\infty} \hat{\lambda}_i \phi_i(x)\phi_i(s)$, where

$$\hat{\lambda}_i = \begin{cases} \lambda_i & \text{if } \lambda_i \geq 0, \\ 0 & \text{if } \lambda_i < 0. \end{cases}$$

Define $T_{K_{\hat{B}}} \colon L^2(X) \to L^2(X)$ to be the Hilbert–Schmidt integral operator associated to the kernel $K_{\hat{B}}$. Note that the eigenfunctions ϕ_i for T_{K_B} (with eigenvalues λ_i) are also the eigenfunctions for $T_{K_{\hat{B}}}$ (with eigenvalues $\hat{\lambda}_i$). By Mercer's Theorem (Theorem 5.2.3), $K_{\hat{B}}$ converges uniformly.

(v) Define the MDS embedding of X into ℓ^2 via the map $f \colon X \to \ell^2$ given by

$$f(x) = \left(\sqrt{\hat{\lambda}_1}\phi_1(x), \sqrt{\hat{\lambda}_2}\phi_2(x), \sqrt{\hat{\lambda}_3}\phi_3(x), \dots \right)$$

for all $x \in X$. We denote the MDS embedding $f(X)$ by $\hat{X}$.

Similarly, define the MDS embedding of X into $\mathbb{R}^m$ via the map $f_m \colon X \to \mathbb{R}^m$ given by

$$f_m(x) = \left(\sqrt{\hat{\lambda}_1}\phi_1(x), \sqrt{\hat{\lambda}_2}\phi_2(x), \dots, \sqrt{\hat{\lambda}_m}\phi_m(x) \right)$$

for all $x \in X$. We denote the image of the MDS embedding by $\hat{X}_m = f_m(X)$.

(vi) Define the measure $\hat{\mu}$ on $(\hat{X}, \hat{d})$ to be the push-forward measure of μ with respect to map f, where $\mathcal{B}(\hat{X})$ is the Borel σ-algebra of $\hat{X}$ with respect to topology induced by $\hat{d}$, the Euclidean metric in $\mathbb{R}^m$ or the ℓ^2 norm. Indeed, the function f is measurable by Corollary 6.3.2 below.

Proposition 6.3.1. *The MDS embedding map $f \colon X \to \ell^2$ defined by*

$$f(x) = \left(\sqrt{\hat{\lambda}_1}\phi_1(x), \sqrt{\hat{\lambda}_2}\phi_2(x), \sqrt{\hat{\lambda}_3}\phi_3(x), \dots \right)$$

is a continuous map.

Proof. Define the sequence of embeddings $g_m \colon X \to \ell^2$, for $m \in \mathbb{N}$, by

$$g_m(x) = \left(\sqrt{\hat{\lambda}_1}\phi_1(x), \dots, \sqrt{\hat{\lambda}_m}\phi_m(x), 0, 0, \dots \right).$$

By Mercer's Theorem (Theorem 5.2.3), we have that the eigenfunctions $\phi_i \colon X \to \mathbb{R}$ are continuous for all $i \in \mathbb{N}$. It follows that g_m is a continuous map for any $m < \infty$.

Consider the sequence of partial sums

$$K_m(x,x) = \sum_{i=1}^{m} \hat{\lambda}_i \phi_i^2(x).$$

By Mercer's Theorem, $K_m(x,x)$ converges uniformly to $K(x,x) = \sum_{i=1}^{\infty} \hat{\lambda}_i \phi_i^2(x)$, as $m \to \infty$. Therefore, for any $\epsilon > 0$, there exists some $N(\epsilon) \in \mathbb{N}$ such that for all $m \geq N(\epsilon)$,

$$\|g_m(x) - f(x)\|_2^2 = \sum_{i=m+1}^{\infty} \hat{\lambda}_i \phi_i^2(x) = |K_m(x,x) - K(x,x)| < \epsilon \text{ for all } x \in X.$$

Therefore, g_m converges uniformly to f as $m \to \infty$. Since the uniform limit of continuous functions is continuous, it follows that $f \colon X \to \ell^2$ is a continuous map. □

With respect to Borel measures every continuous map is measurable, and therefore we immediately obtain the following corollary.

Corollary 6.3.2. *The MDS embedding map $f \colon X \to \ell^2$ is measurable.*

Theorem 6.3.3. *A metric measure space (X, d_X, μ) is Euclidean if and only if T_{K_B} is a positive semi-definite operator on $L^2(X,\mu)$.*

Proof. Suppose (X, d_X, μ) is a Euclidean metric measure space. For any discretization X_n of X and for any $x_i, x_j \in X_n$, the matrix $\mathbf{B}_{ij} = K_B(x_i, x_j)$ is a positive semi-definite matrix by Theorem 4.2.1. Furthermore, by Mercer's Theorem (Theorem 5.2.3) we have that T_{K_B} is a positive semi-definite operator on $L^2(X,\mu)$.

Now, suppose that T_{K_B} is a positive semi-definite operator on $L^2(X,\mu)$. By Mercer's Theorem (Theorem 5.2.3), K_B is a positive semi-definite kernel, and furthermore $K_B(x,s) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(s)$

converges uniformly in $L^2_{\mu\otimes\mu}(X\times X)$. Thus, by Proposition 6.2.1 we have

$$d_X^2(x,s) = K_B(x,x) + K_B(s,s) - 2K_B(x,s) = \sum_{i=1}^{\infty} \lambda_i(\phi_i(x)-\phi_i(s))^2 = \|f(x)-f(s)\|_2^2.$$

Therefore, (X, d_X, μ) is a Euclidean measure metric space.

□

6.4 Strain Minimization

We begin this section by defining a loss function, which in particular is a Strain function. We then show that the MDS method for metric measure spaces described in Section 6.3 minimizes this loss function.

Define the Strain function of f as follows,

$$\mathrm{Strain}(f) = \|T_{K_B} - T_{K_{\hat{B}}}\|_{HS}^2 = \mathrm{Tr}((T_{K_B} - T_{K_{\hat{B}}})^2) = \int\int \left(K_B(x,t) - K_{\hat{B}}(x,t)\right)^2 \mu(dt)\mu(\mathrm{d}x),$$

which is well-defined (see Remark 5.1.15).

In order to show that MDS for metric measure spaces minimizes the Strain, we will need the following two lemmas.

Lemma 6.4.1. *Let $\lambda \in \ell^2$, let $S = \{\hat{\lambda} \in \ell^2 \mid \hat{\lambda}_i \geq 0$ for all $i\}$, and define $\bar{\lambda} \in S$ by*

$$\bar{\lambda}_i = \begin{cases} \lambda_i & \text{if } \lambda_i \geq 0 \\ 0 & \text{if } \lambda_i < 0. \end{cases}$$

Then $\|\lambda - \bar{\lambda}\|_2 \leq \|\lambda - \hat{\lambda}\|_2$ for all $\hat{\lambda} \in S$.

Lemma 6.4.2. *Let $\lambda \in \ell^2$ have sorted entries $\lambda_1 \geq \lambda_2 \geq \ldots$, let $m \geq 0$, let*

$$S_m = \{\hat{\lambda} \in \ell^2 \mid \hat{\lambda}_i \geq 0 \textit{ for all } i \textit{ with at most } m \textit{ entries positive}\},$$

and define $\bar{\lambda} \in S_m$ *by*

$$\bar{\lambda}_i = \begin{cases} \lambda_i & \textit{if } \lambda_i \geq 0 \textit{ and } i \leq m \\ 0 & \textit{if } \lambda_i < 0 \textit{ or } i > m. \end{cases}$$

Then $\|\lambda - \bar{\lambda}\|_2 \leq \|\lambda - \hat{\lambda}\|_2$ *for all* $\hat{\lambda} \in S_m$.

The proofs of these lemmas are straightforward and hence omitted.

The following theorem generalizes [3, Theorem 14.4.2], or equivalently [37, Theorem 2], to the infinite case. Our proof is organized analogously to the argument in [37, Theorem 2].

Theorem 6.4.3. *Let* (X, d, μ) *be a bounded (and possibly non-Euclidean) metric measure space. Then* $\mathrm{Strain}(f)$ *is minimized over all maps* $f\colon X \to \ell^2$ *or* $f\colon X \to \mathbb{R}^m$ *when* f *is the MDS embedding given in Section 6.3.*

Proof. Let $K_B\colon X \times X \to \mathbb{R}$ be defined as in equation (6.3). For simplicity of notation, let $T_B\colon L^2(X,\mu) \to L^2(X,\mu)$ denote the Hilbert–Schmidt integral operator associated to K_B. So

$$[T_B](g)(x) = \int K_B(x,s)g(s)\mu(\mathrm{d}s).$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots$ denote the eigenvalues of T_B, some of which might be negative. By the spectral theorem of compact self-adjoint operators (Theorem 5.2.2), the eigenfunctions $\{\phi_i\}_{i\in\mathbb{N}}$ of T_B form an orthonormal basis of $L^2(X,\mu)$, and the operator T_B can be expressed as

$$T_B = \sum_i \lambda_i \langle \phi_i, \cdot \rangle \phi_i = \sum_i \lambda_i \phi_i \otimes \phi_i.$$

Let $\{e_i\}_{i\in\mathbb{N}}$ be another orthonormal basis of $L^2(X,\mu)$. Define $M_B\colon L^2(X,\mu) \to L^2(X,\mu)$ as follows

$$M_B = \sum_i \langle e_i, \cdot \rangle \phi_i = \sum_i \phi_i \otimes e_i.$$

Indeed, M_B is an orthogonal operator and it can be thought of as a "change of basis" operator from the $\{e_i\}$ basis to the $\{\phi_i\}$ basis. The adjoint of M_B, denoted by $M_B' \colon L^2(X, \mu) \to L^2(X, \mu)$, is defined as follows

$$M_B' = \sum_i \langle \phi_i, \cdot \rangle e_i = \sum_i e_i \otimes \phi_i.$$

Lastly, define the Hilbert–Schmidt operator $S_B \colon L^2(X, \mu) \to L^2(X, \mu)$ as follows

$$S_B = \sum_i \lambda_i \langle e_i, \cdot \rangle e_i = \sum_i \lambda_i e_i \otimes e_i.$$

With respect to the basis $\{e_i\}$, the operator S_B can be thought of as an infinite analogue of a diagonal matrix. It can be shown that $T_B = M_B \circ S_B \circ M_B'$, consequently $M_B' \circ T_B \circ M_B = S_B$ since M_B is an orthogonal operator.

We are attempting to minimize $\mathrm{Strain}(f) = \mathrm{Tr}((T_B - T_{\hat{B}})^2)$ over all symmetric positive semi-definite L^2-kernels $K_{\hat{B}}$ of rank at most m (for $f \colon X \to \mathbb{R}^m$), where we allow $m = \infty$ (for $f \colon X \to \ell^2$). Here $T_{\hat{B}}$ is defined as $[T_{\hat{B}}](g)(x) = \int K_{\hat{B}}(x, s) g(s) \mu(\mathrm{d}s)$. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq 0$ denote the eigenvalues of $T_{\hat{B}}$ with corresponding eigenfunctions $\{\hat{\phi}_i\}_{i \in \mathbb{N}}$, where in the case of $f \colon X \to \mathbb{R}^m$ we require $\hat{\phi}_i = 0$ for $i > m$. We have a similar factorization $T_{\hat{B}} = M_{\hat{B}} \circ S_{\hat{B}} \circ M'_{\hat{B}}$ for the analogously defined operators $M_{\hat{B}}$, $S_{\hat{B}}$, and $M'_{\hat{B}}$. For the time being, we will think of $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq 0$ as fixed, and will optimize $T_{\hat{B}}$ over all possible "change of basis" orthogonal operators $M_{\hat{B}}$.

Note $M_B' \circ T_{\hat{B}} \circ M_B = T_G \circ S_{\hat{B}} \circ T_G'$, where $T_G = M_B' \circ M_{\hat{B}}$ is an orthogonal operator. Therefore, we have

$$
\begin{aligned}
\mathrm{Strain}(f) &= \mathrm{Tr}((T_B - T_{\hat{B}})^2) \\
&= \mathrm{Tr}((T_B - T_{\hat{B}})M_B M_B'(T_B - T_{\hat{B}})) \\
&= \mathrm{Tr}(M_B'(T_B - T_{\hat{B}})(T_B - T_{\hat{B}})M_B) && \text{by Proposition 5.1.11} \\
&= \mathrm{Tr}(M_B'(T_B - T_{\hat{B}})M_B M_B'(T_B - T_{\hat{B}})M_B) \\
&= \mathrm{Tr}((S_B - T_G S_{\hat{B}} T_G')^2) \\
&= \mathrm{Tr}(S_B^2) - \mathrm{Tr}(S_B T_G S_{\hat{B}} T_G') - \mathrm{Tr}(T_G S_{\hat{B}} T_G' S_B) + \mathrm{Tr}(S_{\hat{B}}^2) \\
&= \mathrm{Tr}(S_B^2) - 2\mathrm{Tr}(S_B T_G S_{\hat{B}} T_G') + \mathrm{Tr}(S_{\hat{B}}^2) && (6.4)
\end{aligned}
$$

In the above, we are allowed to apply Proposition 5.1.11 because the fact that $(T_B - T_{\hat{B}})$ is Hilbert–Schmidt implies that $(T_B - T_{\hat{B}})M_B$, and hence also $M_B'(T_B - T_{\hat{B}})$, are Hilbert–Schmidt.

The loss function $\mathrm{Strain}(f)$ is minimized by choosing the orthogonal operator T_G that maximizes $\mathrm{Tr}(S_B T_G S_{\hat{B}} T_G')$. We compute

$$
\mathrm{Tr}(S_B T_G S_{\hat{B}} T_G') = \sum_{i,j} \lambda_i \hat{\lambda}_j \langle \hat{\phi}_j, \phi_i \rangle^2 = \sum_i \lambda_i \left(\sum_j \hat{\lambda}_j \langle \hat{\phi}_j, \phi_i \rangle^2 \right) = \sum_i h_i \lambda_i, \qquad (6.5)
$$

where $h_i = \sum_j \hat{\lambda}_j \langle \hat{\phi}_j, \phi_i \rangle^2$. Notice that

$$
h_i \geq 0 \text{ and } \sum_i h_i = \sum_j \hat{\lambda}_j \sum_i \langle \hat{\phi}_j, \phi_i \rangle^2 = \sum_j \hat{\lambda}_j.
$$

This follows from the fact that $\hat{\phi}_j = \sum_i \langle \hat{\phi}_j, \phi_i \rangle \phi_i$ and

$$
\sum_i \langle \hat{\phi}_j, \phi_i \rangle^2 = \left\langle \hat{\phi}_j, \sum_i \langle \hat{\phi}_j, \phi_i \rangle \phi_i \right\rangle = \langle \hat{\phi}_j, \hat{\phi}_j \rangle = 1.
$$

Since $\sum_i h_i = \sum_j \hat{\lambda}_j$ is fixed and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq 0$, we maximize (6.5) by choosing h_1 as large as possible. Indeed, maximize h_1 by choosing $\hat{\phi}_1 = \phi_1$, so that $\langle \hat{\phi}_1, \phi_1 \rangle = 1$ and $\langle \hat{\phi}_1, \phi_i \rangle = 0$

for all $i \neq 1$. We will show that this choice of $\hat{\phi}_1$ can be completed into an orthonormal basis $\{\hat{\phi}_j\}_{j\in\mathbb{N}}$ for $L^2(X)$ in order to form a well-defined and optimal positive semidefinite kernel $K_{\bar{B}}$ (of rank at most m in the $f\colon X \to \mathbb{R}^m$ case). Next, note $\sum_{i=2} h_i\lambda_i$ is maximized by choosing h_2 as large as possible, which is done by choosing $\hat{\phi}_2 = \phi_2$. It follows that for $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq 0$ fixed, (6.5) is maximized, and hence (6.4) is minimized, by choosing $\hat{\phi}_i = \phi_i$ for all i. Hence,

$$T_G = \sum_{i,j} \langle \hat{\phi}_i, \phi_j \rangle e_j \otimes e_i = \sum_i e_i \otimes e_i.$$

Therefore, we can do no better than choosing T_G to be the identity operator, giving $M_{\hat{B}} = M_B$.

We will now show how to choose the eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq 0$. As in Lemmas 6.4.1 and 6.4.2, for $f\colon X \to \ell^2$ we define

$$\bar{\lambda}_i = \begin{cases} \lambda_i & \text{if } \lambda_i \geq 0 \\ 0 & \text{if } \lambda_i < 0, \end{cases}$$

and for $f\colon X \to \mathbb{R}^m$ we define

$$\bar{\lambda}_i = \begin{cases} \lambda_i & \text{if } \lambda_i \geq 0 \text{ and } i \leq m \\ 0 & \text{if } \lambda_i < 0 \text{ or } i > m. \end{cases}$$

Define $K_{\bar{B}}(x,s) = \sum_{i=1}^{\infty} \bar{\lambda}_i \phi_i(x)\phi_i(s)$, where each eigenfunction ϕ_i of $K_{\bar{B}}$ is the eigenfunction ϕ_i of K_B corresponding to the eigenvalue $\bar{\lambda}_i$. We compute, that over all possible choices of eigenvalues, the optimal Strain is given by the choices $\bar{\lambda}_i$:

$$
\begin{aligned}
\mathrm{Strain}(f) &= \mathrm{Tr}((T_B - T_{\hat{B}})^2) \\
&= \mathrm{Tr}((S_B - T_G S_{\hat{B}} T_G')^2) \\
&\geq \mathrm{Tr}((S_B - S_{\hat{B}})^2) \\
&= \|\lambda - \hat{\lambda}\|_2^2 \\
&\geq \|\lambda - \bar{\lambda}\|_2^2 && \text{by Lemma 6.4.1 or 6.4.2} \\
&= \mathrm{Tr}((S_B - S_{\bar{B}})^2) \\
&= \mathrm{Tr}((M_B(S_B - S_{\bar{B}})M_B')^2) && \text{by Proposition 5.1.11} \\
&= \mathrm{Tr}((T_B - T_{\bar{B}})^2).
\end{aligned}
$$

Therefore, the loss function $\mathrm{Strain}(f)$ is minimized when f and $T_{\hat{B}}$ are defined via the MDS embedding in Section 6.3. □

The following table shows a comparison of various elements of classical MDS and infinite MDS (as described in section 6.3). Our table is constructed analogously to a table on the Wikipedia page [40] that shows a comparison of various elements of Principal Component Analysis (PCA) and Functional Principal Component Analysis (FPCA). In Chapter 8, we address convergence questions for MDS more generally.

Table 6.1: A comparison table of various elements of classical MDS and infinite MDS.

Elements	**Classical MDS**	**Infinite MDS**
Data	(X_n, d)	(X, d_X, μ)
Distance Representation	$D_{i,j} = d(x_i, x_j), \quad D \in \mathcal{M}_{n\times n}$	$K_D(x,s) = d_X(x,s) \in L^2_{\mu\otimes\mu}(X \times X)$
Linear Operator	$B = -\frac{1}{2} H D^{(2)} H$	$[T_{K_B}\phi](x) = \int\limits_X K_B(x,s)\phi(s)\mu(\mathrm{d}s)$
Eigenvalues	$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$	$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$
Eigenvectors/Eigenfunctions	$v^{(1)}, v^{(2)}, \ldots, v^{(m)} \in \mathbb{R}^n$	$\phi_1(x), \phi_2(x), \ldots \in L^2(X)$
Embedding in $\mathbb{R}^m$ or ℓ^2	$f(x_i) = \left(\sqrt{\lambda_1} v_1^{(i)}, \sqrt{\lambda_2} v_2^{(i)}, \ldots, \sqrt{\lambda_m} v_m^{(i)}\right)$	$f(x) = \left(\sqrt{\hat{\lambda}_1}\phi_1(x), \sqrt{\hat{\lambda}_2}\phi_2(x), \sqrt{\hat{\lambda}_3}\phi_3(x), \ldots\right)$
Strain Minimization	$\sum\limits_{i,j=1}^{n} (b_{i,j} - \hat{b}_{i,j})^2$	$\int\int \left(K_B(x,t) - K_{\hat{B}}(x,t)\right)^2 \mu(\mathrm{d}t)\mu(\mathrm{d}x)$

Chapter 7

MDS of the Circle

In this chapter, we consider the MDS embeddings of the circle equipped with the (non-Euclidean) geodesic metric. The material in this section is closely related to [39], even though [39] was written prior to the invention of MDS. By using the known eigendecomposition of circulant matrices, we are able to give an explicit description of the MDS embeddings of evenly-spaced points from the circle. This is a motivating example for the convergence properties studied in Chapter 8, which will show that the MDS embeddings of more and more evenly-spaced points will converge to the MDS embedding of the entire geodesic circle. We also remark that the geodesic circle is a natural example of a metric space whose MDS embedding in ℓ^2 is better (in the sense of Strain-minimization) than its MDS embedding into $\mathbb{R}^m$ for any finite m.

We describe the eigenvalues and eigenvectors of circulant matrices in Section 7.1, and use this to describe the MDS embeddings of evenly-spaced points from the circle in Section 7.2. In Section 7.3, we describe the relationship between the MDS embedding of the circle and the much earlier work of [39, 42]. We also refer the reader to the conclusion (Chapter 9) for open questions on the MDS embeddings of geodesic n-spheres S^n into $\mathbb{R}^m$.

7.1 Background on Circulant Matrices

Let $\mathbf{B}$ be an $n \times n$ matrix. The matrix $\mathbf{B}$ is *circulant* if each row is a cyclic permutation of the first row, in the form as shown below.

$$\mathbf{B} = \begin{pmatrix} b_0 & b_1 & b_2 & \dots & b_{n-3} & b_{n-2} & b_{n-1} \\ b_{n-1} & b_0 & b_1 & \dots & b_{n-4} & b_{n-3} & b_{n-2} \\ b_{n-2} & b_{n-1} & b_0 & \dots & b_{n-5} & b_{n-4} & b_{n-3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ b_3 & b_4 & b_5 & \dots & b_0 & b_1 & b_2 \\ b_2 & b_3 & b_4 & \dots & b_{n-1} & b_0 & b_1 \\ b_1 & b_2 & b_3 & \dots & b_{n-2} & b_{n-1} & b_0 \end{pmatrix}$$

The first row of a circulant matrix determines the rest of the matrix. Furthermore, a circulant matrix $\mathbf{B}$ has a basis of eigenvectors of the form $x_k(n) = \begin{pmatrix} w_n^{0k} & w_n^{1k} & \dots & w_n^{(n-1)k} \end{pmatrix}^\top$ for $0 \leq k \leq n-1$, where $w_n = e^{\frac{2\pi i}{n}}$ and $\top$ denotes the transpose of a matrix. The eigenvalue corresponding to $x_k(n)$ is

$$\lambda_k(n) = \sum_{j=0}^{n-1} b_j w_n^{jk} = b_0 w_n^{0k} + b_1 w_n^{1k} + \cdots + b_{n-1} w_n^{(n-1)k}. \tag{7.1}$$

If the circulant matrix $\mathbf{B}$ is also symmetric, then $b_i = b_{n-i}$ for $1 \leq i \leq n-1$. Thus, $\mathbf{B}$ has the form

$$\mathbf{B} = \begin{pmatrix} b_0 & b_1 & b_2 & \dots & b_3 & b_2 & b_1 \\ b_1 & b_0 & b_1 & \dots & b_4 & b_3 & b_2 \\ b_2 & b_1 & b_0 & \dots & b_5 & b_4 & b_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ b_3 & b_4 & b_5 & \dots & b_0 & b_1 & b_2 \\ b_2 & b_3 & b_4 & \dots & b_1 & b_0 & b_1 \\ b_1 & b_2 & b_3 & \dots & b_2 & b_1 & b_0 \end{pmatrix}. \tag{7.2}$$

The eigenvalues $\lambda_k(n)$ are then real and of the form

$$\lambda_k = b_0 + 2b_1\Re(w_n^{1k}) + \cdots + 2b_{\frac{n-1}{2}}\Re(w_n^{\frac{(n-1)}{2}k}) \qquad \text{if } n \text{ is odd, and}$$

$$\lambda_k = b_0 + b_{\frac{n}{2}}w_n^{\frac{n}{2}k} + 2b_1\Re(w_n^{1k}) + \cdots + 2b_{\frac{n}{2}-1}\Re(w_n^{(\frac{n}{2}-1)k}) \qquad \text{if } n \text{ is even.}$$

7.2 MDS of Evenly-Spaced Points on the Circle

Let S^1 be the unit circle (i.e. with circumference 2π), equipped with the geodesic metric d which can be simply thought of as the shortest path between two given points in a curved space (in this case, the circle). We let S_n^1 denote a set of n evenly spaced points on S^1. In Proposition 7.2.6, we show that the MDS embedding of S_n^1 in $\mathbb{R}^m$ lies, up to a rigid motion, on the curve $\gamma_n \colon S^1 \to \mathbb{R}^m$ defined by

$$\gamma_m(\theta) = (a_1(n)\cos(\theta), a_1(n)\sin(\theta), a_3(n)\cos(3\theta), a_3(n)\sin(3\theta), a_5(n)\cos(5\theta), a_5(n)\sin(5\theta), \ldots) \in \mathbb{R}^m,$$

where $\lim_{n\to\infty} a_j(n) = \frac{\sqrt{2}}{j}$ (with j odd).

Let $\mathbf{D}$ be the distance matrix of S_n^1, which is determined (up to symmetries of the circle) by

$$\frac{n}{2\pi}d_{0j} = \begin{cases} j & \text{for } 0 \le j \le \lfloor \frac{n}{2} \rfloor \\ n-j & \text{for } \lceil \frac{n}{2} \rceil \le j \le n-1, \end{cases}$$

where $\lfloor . \rfloor$ denotes the floor function and $\lceil . \rceil$ denotes the ceiling function. Let $\mathbf{A} = (a_{ij})$ with $a_{ij} = -\frac{1}{2}d_{ij}^2$ and let $\mathbf{B} = \mathbf{HAH}$, where $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. The distance matrix $\mathbf{D}$ is real-symmetric circulant, and it follows that $\mathbf{B}$ is real-symmetric circulant with its form as shown in equation (7.2). After applying symmetries of the circle, the entries of the first row vector $(b_0, b_1, \cdots, b_{n-1})$ of $\mathbf{B}$ can be written explicitly as

$$b_j = -\frac{1}{2}\left(d_{0j}^2 - \frac{1}{n}\sum_{k=0}^{n-1} d_{0k}^2 - \frac{1}{n}\sum_{k=0}^{n-1} d_{kj}^2 + \frac{1}{n^2}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1} d_{kl}^2\right) = -\frac{1}{2}(d_{0j}^2 - c_n),$$

where a formula for $c_n = \frac{1}{n}\sum_{k=0}^{n-1} d_{0k}^2$ can be computed explicitly[1]. Furthermore, let $\lambda_k(n)$ denote the kth eigenvalue of the matrix $\mathbf{B}$ corresponding to the kth eigenvector $x_k(n)$. A basis of eigenvectors for $\mathbf{B}$ consists of $x_k(n) = \begin{pmatrix} w_n^{0k} & w_n^{1k} & \dots & w_n^{(n-1)k} \end{pmatrix}^\top$ for $0 \le k \le n-1$, where $w_n = e^{\frac{i2\pi}{n}}$.

Lemma 7.2.1. *We have* $\lambda_k(n) = -\frac{1}{2}\sum_{j=0}^{n-1} d_{0j}^2 w_n^{jk}$ *for* $0 \le k \le n-1$.

Proof. We compute

$$\lambda_k(n) = \sum_{j=0}^{n-1} b_j w_n^{jk} = \sum_{j=0}^{n-1} -\frac{1}{2}(d_{0j}^2 - c_n) w_n^{jk} = -\frac{1}{2}\sum_{j=0}^{n-1} d_{0j}^2 w_n^{jk} + \frac{1}{2}c_n \sum_{j=0}^{n-1} w_n^{jk},$$

and the property follows by noting that $\sum_{j=0}^{n-1} w_n^{jk} = 0$. □

Corollary 7.2.2. *The* k*th eigenvalue* $\lambda_k(n)$ *corresponding to the eigenvector* $x_k(n)$ *satisfies*

$$\lambda_0(n) = 0, \quad \textit{and} \quad \lim_{n\to\infty} \frac{\lambda_k(n)}{n} = \frac{(-1)^{k+1}}{k^2} \quad \textit{for} \quad k \ge 1.$$

Proof. If $k = 0$, we have $x_0(n) = \mathbf{1}$ with eigenvalue $\lambda_0(n) = 0$ since $\mathbf{B}$ is a double-centered matrix. Hence we restrict attention to $1 \le k \le n-1$. By Lemma 7.2.1, we have we have

$$\frac{\lambda_k(n)}{n} = -\frac{1}{2n}\sum_{j=0}^{n-1} d_{0j}^2 w_n^{jk} = -\frac{1}{2n}\left(\frac{2\pi}{n}\right)^2 \sum_{j=-\lfloor\frac{n}{2}\rfloor}^{\lfloor\frac{n-1}{2}\rfloor} j^2 e^{(\frac{2\pi}{n}jk)i} = -2\pi^2\left(\frac{1}{n}\right)\sum_{j=-\lfloor\frac{n}{2}\rfloor}^{\lfloor\frac{n-1}{2}\rfloor}\left(\frac{j}{n}\right)^2 e^{(2\pi(\frac{j}{n})k)i} =: S_n.$$

Since S_n is the left-hand Riemann sum (with n subintervals) of the below integral, we use integration by parts to get

[1] If n is odd we have $c_n = \frac{2}{n}\sum_{k=0}^{\frac{n-1}{2}}\left(\frac{2\pi k}{n}\right)^2 = \frac{\pi^2}{3n^2}(n^2-1)$, and if n is even we have

$$c_n = \frac{1}{n}\left(d_{0,\frac{n}{2}}^2 + 2\sum_{k=0}^{\frac{n}{2}-1} d_{0k}^2\right) = \frac{1}{n}\left(\pi^2 + 2\sum_{k=0}^{\frac{n}{2}-1}\left(\frac{2\pi k}{n}\right)^2\right) = \frac{\pi^2}{3n^2}(n-1)(n-2) + \frac{\pi^2}{n}.$$

$$\lim_{n\to\infty}\frac{\lambda_k(n)}{n}=\lim_{n\to\infty}S_n=-2\pi^2\int_{-\frac{1}{2}}^{\frac{1}{2}}x^2e^{2\pi xki}\ \mathrm{d}x=\frac{(-1)^{k+1}}{k^2}. \tag{7.3}$$

□

Lemma 7.2.3. *For all k odd, there exists $N\in\mathbb{N}$ sufficiently large (possibly depending on p odd) such that for all $n\geq N$, the eigenvalues $\lambda_k(n)$ satisfy the following property*

$$\lambda_1(n)\geq\lambda_3(n)\geq\lambda_5(n)\geq\cdots\geq\lambda_p(n)\geq 0,$$

where $\lambda_k(n)$ is the eigenvalue corresponding to $x_k(n)$.

Proof. From equation (7.3), for k odd we have,

$$\lim_{n\to\infty}\frac{\lambda_k(n)}{n}=\frac{1}{k^2}.$$

Therefore, for each k odd, $\exists N_p\in\mathbb{N}$ such that $\forall n\geq N_p$, we have

$$\lambda_1(n)\geq\lambda_3(n)\geq\lambda_5(n)\geq\cdots\geq\lambda_p(n)\geq 0. \tag{7.4}$$

□

Remark 7.2.4. We conjecture that Lemma 7.2.3 is true for $N_p=1$.

For a real-symmetric circulant matrix, the real and imaginary parts of the eigenvectors $x_k(n)$ are eigenvectors with the same eigenvalue. These eigenvectors correspond to a discrete cosine transform and a discrete sine transform. Let $u_k(n)$ and $v_k(n)$ denote the real and imaginary parts of $x_k(n)$ respectively. In general,

$$u_k(n)=\begin{pmatrix}1 & \cos\theta & \cos 2\theta & \dots & \cos(n-1)\theta\end{pmatrix}^\top,$$

and

$$v_k(n) = \begin{pmatrix} 0 & \sin\theta & \sin 2\theta & \dots & \sin(n-1)\theta \end{pmatrix}^\top,$$

where $\theta = \frac{2\pi k}{n}$.

Since $\mathbf{B}$ is a real symmetric matrix, its orthogonal eigenvectors can also be chosen real. We have $\Re(w_n^k) = \Re(w_n^{n-k})$ and $\Im(w_n^k) = -\Im(w_n^{n-k})$. A new basis of n eigenvectors in $\mathbb{R}^n$ of $\mathbf{B}$ can be formed from $u_k(n)$ and $v_k(n)$ as follows. Let $\mathcal{E} = \{u_k(n), v_k(n) \mid k = 1, 2, \cdots, \frac{n}{2} - 1\}$ and $\mathcal{O} = \{u_k(n), v_k(n) \mid k = 1, 2, \cdots, \frac{n-1}{2}\}$. If n is even, a set of n linearly independent eigenvectors is

$$\mathcal{B}_e = \mathcal{E} \cup \{u_0(n), u_{\frac{n}{2}}(n)\}.$$

Furthermore, if n is odd, a set of n linearly independent eigenvectors is

$$\mathcal{B}_o = \mathcal{O} \cup \{u_0(n)\}.$$

In each of the sets $\mathcal{E}$ and $\mathcal{O}$, the eigenvalue of $u_k(n)$ is the same as the eigenvalue of $v_k(n)$ for each k. So, the eigenvalues in $\mathcal{E}$ and $\mathcal{O}$ come in pairs. The only eigenvalues that don't come in pairs correspond to eigenvectors $x_k(n)$ that are purely real, namely $u_0(n)$, and $u_{\frac{n}{2}}(n)$ (for n even). Indeed, the eigenvalue corresponding to $u_0(n)$ is $\lambda_0(n) = 0$. Furthermore, the new eigenvectors are real and mutually orthogonal.

Lemma 7.2.5. *The ℓ^2-norm of $u_k(n)$ and $v_k(n)$ satisfy the following property*

$$\lim_{n\to\infty} \frac{\|u_k(n)\|}{\sqrt{n}} = \lim_{n\to\infty} \frac{\|v_k(n)\|}{\sqrt{n}} = \frac{1}{\sqrt{2}}. \tag{7.5}$$

Proof. Consider the following where $\theta = \frac{2\pi k}{n}$, and where we have used the fact that a limit of Riemann sums converges to the corresponding Riemann integral.

$$\lim_{n\to\infty}\frac{\|u_k(n)\|^2}{n}=\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\cos^2(j\theta)=\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\cos^2(2\pi k(\tfrac{j}{n}))=\int_0^1\cos^2(2\pi kx)\,\mathrm{d}x=\frac{1}{2},\text{ and}$$

$$\lim_{n\to\infty}\frac{\|v_k(n)\|^2}{n}=\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\sin^2(j\theta)=\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\sin^2(2\pi k(\tfrac{j}{n}))=\int_0^1\sin^2(2\pi kx)\,\mathrm{d}x=\frac{1}{2}.$$

The definite integrals can be computed by hand using double angle formulas. Therefore, the result follows. □

Let $\mathbf{\Lambda}_m$ be the $m\times m$ diagonal matrix of the largest m eigenvalues of $\mathbf{B}$ sorted in descending order, and let $\mathbf{\Gamma}_m$ be the $n\times m$ matrix of the corresponding m eigenvectors. From Lemma 7.2.3, $\mathbf{\Lambda}_m$ can be written as follows,

$$\mathbf{\Lambda}_m=\begin{pmatrix}\lambda_1(n)&&&&&&\\&\lambda_1(n)&&&&&\\&&\lambda_3(n)&&&&\\&&&\lambda_3(n)&&&\\&&&&\ddots&&\\&&&&&\lambda_m(n)\end{pmatrix}$$

and $\mathbf{\Gamma}_m$ can be constructed in different of ways. We'll construct $\mathbf{\Gamma}_m$ as follows,

$$\mathbf{\Gamma}_m=\begin{pmatrix}\frac{u_1(n)}{\|u_1(n)\|}&\frac{v_1(n)}{\|v_1(n)\|}&\frac{u_3(n)}{\|u_3(n)\|}&\frac{v_3(n)}{\|v_3(n)\|}&\cdots&\frac{v_m(n)}{\|v_m(n)\|}\end{pmatrix}.$$

The classical MDS embedding of S^1_n consists of the n points in $\mathbb{R}^m$ whose coordinates are given by the n rows of the $n\times m$ matrix $\mathbf{X}=\mathbf{\Gamma}_m\mathbf{\Lambda}_m^{1/2}$.

Proposition 7.2.6. *The classical MDS embedding of S^1_n lies, up to a rigid motion of $\mathbb{R}^m$, on the curve $\gamma_m\colon S^1\to\mathbb{R}^m$ defined by*

$$\gamma_m(\theta) = (a_1(n)\cos(\theta), a_1(n)\sin(\theta), a_3(n)\cos(3\theta), a_3(n)\sin(3\theta), a_5(n)\cos(5\theta), a_5(n)\sin(5\theta), \ldots) \in \mathbb{R}^m,$$

where $\lim_{n\to\infty} a_j(n) = \frac{\sqrt{2}}{j}$ *(with* j *odd).*

Proof. The coordinates of the points of the MDS embedding of S^1_n are given by the $n \times m$ matrix $\mathbf{X} = \mathbf{\Gamma}_m \mathbf{\Lambda}_m^{1/2}$. This implies the coordinates of the n configuration points in $\mathbb{R}^m$ are given by

$$(a_1(n)\cos(\theta), b_1(n)\sin(\theta), a_3(n)\cos(3\theta), b_3(n)\sin(3\theta), a_5(n)\cos(5\theta), b_5(n)\sin(5\theta), \ldots) \in \mathbb{R}^m,$$

where $\theta = \frac{2\pi k}{n}$ and $0 \leq k \leq n-1$ and (for j odd)

$$a_j(n) = \frac{\sqrt{\lambda_j(n)}}{\|u_j(n)\|} \quad \text{and} \quad b_j(n) = \frac{\sqrt{\lambda_j(n)}}{\|v_j(n)\|}.$$

From Corollary 7.2.2 and Lemma 7.2.5, we have

$$\lim_{n\to\infty} a_j(n) = \lim_{n\to\infty} \frac{\sqrt{\lambda_j(n)}}{\|u_j(n)\|} = \lim_{n\to\infty} \frac{\frac{\sqrt{\lambda_j(n)}}{\sqrt{n}}}{\frac{\|u_j(n)\|}{\sqrt{n}}} = \frac{\sqrt{2}}{j},$$

and similarly $\lim_{n\to\infty} b_j(n) = \frac{\sqrt{2}}{j}$. Therefore, we can say that the MDS embedding of S^1_n lies, up to a rigid motion of $\mathbb{R}^m$, on the curve $\gamma_m \colon S^1 \to \mathbb{R}^m$ defined by

$$\gamma_m(\theta) = (a_1(n)\cos(\theta), a_1(n)\sin(\theta), a_3(n)\cos(3\theta), a_3(n)\sin(3\theta), a_5(n)\cos(5\theta), a_5(n)\sin(5\theta), \ldots) \in \mathbb{R}^m,$$

where $\lim_{n\to\infty} a_j(n) = \frac{\sqrt{2}}{j}$ (with j odd). □

Indeed, Figure 7.1 shows the MDS configuration in $\mathbb{R}^3$ of 1000 points on S^1 obtained using the three largest positive eigenvalues.

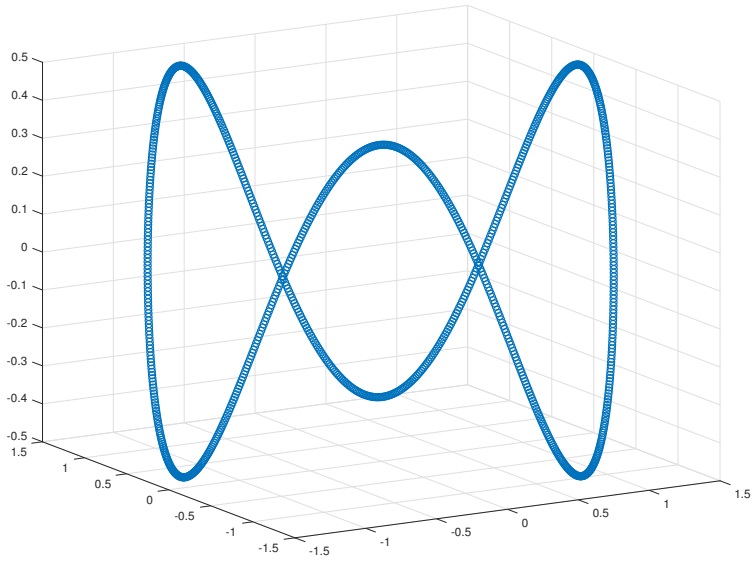


Figure 7.1: MDS embedding of S^1_{1000}.

7.3 Relation to Work of von Neumann and Schoenberg

The MDS embeddings of the geodesic circle are closely related to [39], which was written prior to the invention of MDS. In [39, Theorem 1], von Neumann and Schoenberg describe (roughly speaking) which metrics on the circle one can isometrically embed into the Hilbert space ℓ^2. The geodesic metric on the circle is not one of these metrics. However, the MDS embedding of the geodesic circle into ℓ^2 must produce a metric on S^1 which is of the form described in [39, Theorem 1]. See also [42, Section 5] and [4, 7, 11].

Chapter 8

Convergence of MDS

We saw in the prior chapter how sampling more and more evenly-spaced points from the geodesic circle allowed one to get a sense of how MDS behaves on the entire circle. In this chapter, we address convergence questions for MDS more generally. Convergence is well-understood when each metric space has the same finite number of points [31], but we are also interested in convergence when the number of points varies and is possibly infinite.

This chapter is organized as follows. In Section 8.1, we survey Sibson's perturbation analysis [31] for MDS on a fixed number of n points. Next, in Section 8.2, we survey results of [2, 16] on the convergence of MDS when n points $\{x_1, \ldots, x_n\}$ are sampled from a metric space according to a probability measure μ, in the limit as $n \to \infty$. Unsurprisingly, these results rely on the law of large numbers. In Section 8.3, we reprove these results under the (simpler) deterministic setting when points are not randomly chosen, and instead we assume that the corresponding finite measures $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ (determined by n points) converge to μ. This allows us, in Section 8.4, to consider the more general setting where we have convergence of *arbitrary* probability measures $\mu_n \to \mu$. For example, in what sense do we still have convergence of MDS when each measure μ_n in the converging sequence $\mu_n \to \mu$ has infinite support? Finally, in Section 8.5, we ask about the even more general setting where we have the convergence of arbitrary metric measure spaces $(X_n, d_n, \mu_n) \to (X, d, \mu)$ in the Gromov–Wasserstein distance.

8.1 Robustness of MDS with Respect to Perturbations

In a series of papers [30–32], Sibson and his collaborators consider the robustness of multidimensional scaling with respect to perturbations of the underlying distance or dissimilarity matrix as illustrated in Figure 8.1. In particular, [31] gives quantitative control over the perturbation of

the eigenvalues and vectors determining an MDS embedding in terms of the perturbations of the dissimilarities. These results build upon the fact that if λ and v are a (simple, i.e. non-repeated) eigenvalue and eigenvector of an $n \times n$ matrix B, then one can control the change in λ and v upon a small symmetric perturbation of the entries in B.

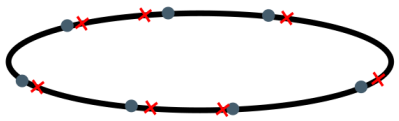


Figure 8.1: Perturbation of the given dissimilarities.

Sibson's perturbation analysis shows that if one is has a converging sequence of $n \times n$ dissimilarity matrices, then the corresponding MDS embeddings of n points into Euclidean space also converge. In the following sections, we consider the convergence of MDS when the number of points is not fixed. Indeed, we consider the convergence of MDS when the number of points is finite but tending to infinity, and alternatively also when the number of points is infinite at each stage in a converging sequence of metric measure spaces.

8.2 Convergence of MDS by the Law of Large Numbers

In this section, we survey results of [2, 16] on the convergence of MDS when more and more points are sampled from a metric space.

Suppose we are given the data set $X_n = \{x_1, \ldots, x_n\}$ with $x_i \in \mathbb{R}^k$ sampled independent and identically distributed (i.i.d.) from an unknown probability measure μ on X. Define $D_{ij} = d(x_i, x_j)$ and the corresponding data-dependent kernel K_n as follows

$$K_n(x,y) = -\frac{1}{2}\left(d^2(x,y) - \frac{1}{n}\sum_{i=1}^{n} d^2(x_i,y) - \frac{1}{n}\sum_{i=1}^{n} d^2(x,x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} d^2(x_i,x_j) \right). \quad (8.1)$$

Define the Gram matrix $M = -\frac{1}{2}HD^{(2)}H$, where $H = I - n^{-1}\mathbf{1}\mathbf{1}^\top$, and note that $M_{ij} = K_n(x_i, x_j)$ for $i, j = 1, \ldots, n$. Assume that the (possibly data-dependent) kernel K_n is bounded (i.e. $|K_n(x, y)| < c$ for all x, y in $\mathbb{R}^k$). We will assume that K_n converges uniformly in its arguments and in probability to its limit K as $n \to \infty$. This means that for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(\sup_{x,y\in\mathbb{R}^k} |K_n(x, y - K(x, y)| \geq \epsilon) = 0.$$

Furthermore, assume K_n is an L^2-kernel. Then associated to K_n is an operator $T_{K_n} \colon L^2(X) \to L^2(X)$ defined as

$$[T_{K_n} f](x) = \frac{1}{n}\sum_{i=1}^{n} K_n(x, x_i) f(x_i).$$

Define $T_K \colon L^2(X) \to L^2(X)$ as

$$[T_K f](x) = \int K(x, s) f(s) \mu(\mathrm{d}s),$$

where

$$K(x, y) = \frac{1}{2}\left(-d(x, y)^2 + \int_X d(w, y)^2 \mu(\mathrm{d}w) + \int_X d(x, z)^2 \mu(\mathrm{d}z) - \int_{X\times X} d(w, z)^2 \mu(\mathrm{d}w \times \mathrm{d}z)\right)$$

is defined as in Section 6.3. Therefore, we obtain the following eigensystems,

$$T_K f_k = \lambda_k f_k \quad \text{and} \quad T_{K_n}\phi_{k,n} = \lambda_{k,n}\phi_{k,n},$$

where (λ_k, ϕ_k) and $(\lambda_{k,n}, \phi_{k,n})$ are the corresponding eigenvalues and eigenfunctions of T_K and T_{K_n} respectively. Furthermore, when we evaluate T_{K_n} at the data points $x_i \in X_n$, we obtain the following eigensystem for M,

$$M v_k = \ell_k v_k,$$

where (ℓ_k, v_k) are the corresponding eigenvalues and eigenvectors.

Lemma 8.2.1. *[2, Proposition 1] T_{K_n} has in its image $m \leq n$ eigenfunctions of the form*

$$\phi_{k,n}(x) = \frac{\sqrt{n}}{\ell_k} \sum_{i=1}^{n} v_k^{(i)} K_n(x, x_i)$$

with corresponding eigenvalues $\lambda_{k,n} = \frac{\ell_k}{n}$ where $v_k^{(i)}$ denotes the ith entry of the kth eigenvector of M associated with the eigenvalue ℓ_k. For $x_i \in X_n$, these functions coincide with the corresponding eigenvectors, $\phi_{k,n}(x_i) = \sqrt{n} v_k^{(i)}$.

Indeed, M has n eigenvalues whereas T_{K_n} has infinitely many eigenvalues, this means that T_{K_n} has at most n nonzero eigenvalues, and that 0 is an eigenvalue of T_{K_n} with infinite multiplicity. In order to compare the finitely many eigenvalues of M with the infinite sequence of eigenvalues of T_{K_n}, some procedure has to be constructed [16].

Suppose that the eigenvalues are all non-negative and sorted in non-increasing order, repeated according to their multiplicity. Thus, we obtain eigenvalue tuples and sequences

$$\lambda(M) = (l_1, \ldots, l_n),$$

where $l_1 \geq \ldots \geq l_n$, and

$$\lambda(T_{K_n}) = (\lambda_1, \lambda_2, \ldots)$$

where $\lambda_1 \geq \lambda_2 \geq \ldots$.

To compare the eigenvalues, first embed $\lambda(M)$ into ℓ^1 by padding the length-n vector with zeroes, obtaining

$$\lambda(M) = (l_1, \ldots, l_n, 0, 0, \ldots).$$

Definition 8.2.2. The *ℓ^2-rearrangement distance* between (countably) infinite sequences x and y is defined as

$$\delta_2(x, y) = \inf_{\pi \in \mathcal{G}(\mathbb{N})} \sum_{i=1}^{\infty} (x_i - y_{\pi(i)})^2$$

where $\mathcal{G}(\mathbb{N})$ is the set of all bijections on $\mathbb{N}$.

In this section and the following sections, the eigenvalues are always ordered in non-increasing order by the spectral theorem of self-adjoint operators. We note that the ℓ^2-rearrangement distance is simply the ℓ^2-distance when the eigenvalues are ordered.

Theorem 8.2.3. *[16, Theorem 3.1] The ordered spectrum of T_{K_n} converges to the ordered spectrum of T_K as $n \to \infty$ with respect to the ℓ^2-distance, namely*

$$\ell^2(\lambda(T_{K_n}), \lambda(T_K)) \to 0 \quad \textit{a.s.}$$

The theorem stated above is in fact only one part, namely equation (3.13), in the proof of [16, Theorem 3.1]. We also remark that [16] uses fairly different notation for the various operators than what we have used here.

Theorem 8.2.4. *[2, Proposition 2] If K_n converges uniformly in its arguments and in probability, with the eigendecomposition of the Gram matrix converging, and if the eigenfunctions $\phi_{k,n}(x)$ of T_{K_n} associated with non-zero eigenvalues converge uniformly in probability, then their limit are the corresponding eigenfunctions of T_K.*

Under the given hypotheses, we are able to formulate a specific set of eigenvalues and eigenfunctions T_K. As illustrated below, the choice of the eigenfunctions $\phi_{k,n}$ of T_{K_n} was made to extend the finite MDS embedding to the infinite MDS embedding described in Section 6.3. However, there are other possible choices of eigenfunctions $\phi_{k,n}$.

Consider the infinite MDS map $f_m \colon X \to \mathbb{R}^m$ defined in Section 6.3 as

$$f_m(x) = \left(\sqrt{\lambda_1}\phi_1(x), \sqrt{\lambda_2}\phi_2(x), \ldots, \sqrt{\lambda_m}\phi_m(x)\right)$$

for all $x \in X$, with kernel $K = K_B$ and the associated operator $T_K = T_{K_B}$ (with eigensystem (λ_k, ϕ_k)). Evaluating $f_m(x)$ at $x_i \in X_n$, we obtain the following finite embedding:

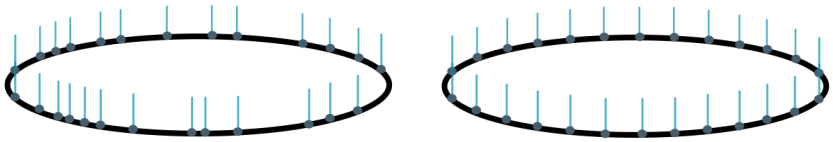
$$\begin{aligned}
f_m(x_i) &= \left(\sqrt{\lambda_1}\phi_1(x_i), \sqrt{\lambda_2}\phi_2(x_i), \ldots, \sqrt{\lambda_m}\phi_m(x_i)\right) \\
&= \left((\lim_{n\to\infty} \sqrt{\lambda_{1,n}} \cdot \phi_{1,n})(x_i), (\lim_{n\to\infty} \sqrt{\lambda_{2,n}} \cdot \phi_{2,n})(x_i), \ldots, (\lim_{n\to\infty} \sqrt{\lambda_{m,n}} \cdot \phi_{m,n})(x_i)\right) \\
&= \left(\left(\lim_{n\to\infty} \sqrt{\frac{\ell_1}{n}} \cdot \sqrt{n} v_1^{(i)}\right), \left(\lim_{n\to\infty} \sqrt{\frac{\ell_2}{n}} \cdot \sqrt{n} v_2^{(i)}\right), \ldots, \left(\lim_{n\to\infty} \sqrt{\frac{\ell_m}{n}} \cdot \sqrt{n} v_m^{(i)}\right)\right) \\
&= \left(\lim_{n\to\infty} (\sqrt{\ell_1} v_1^{(i)}), \lim_{n\to\infty} (\sqrt{\ell_2} v_2^{(i)}), \ldots, \lim_{n\to\infty} (\sqrt{\ell_m} v_m^{(i)})\right),
\end{aligned}$$

which is the finite MDS embedding of X_n into $\mathbb{R}^m$, where the eigensystem of the $n \times n$ inner-product matrix B is denoted by $(\ell_k(n), v_k(n))$. The second equality above is from Theorem 8.2.4, and the third equality above is from Lemma 8.2.1.

8.3 Convergence of MDS for Finite Measures

Though we gave no proofs in the above section, we do so now in the simpler deterministic case when points are not drawn from X at random, but instead we assume that $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is sequence of measures with the support of μ_n consisting of n points $\{x_1, \ldots, x_n\} \subseteq X$, and we assume that μ_n converges to some underlying probability distribution μ on X. Our reason for working deterministically instead of randomly here is so that in Section 8.4, we may consider the convergence of MDS in the more general setting when $\mu_n \to \mu$ are arbitrary probability measures; for example each μ_n may have infinite support.

Figure 8.2a illustrates the case when the points are sampled i.i.d as discussed Section 8.2, in contrast to Figure 8.2b which illustrates the case when the points are sampled in a manner that guarantees convergence of the measures μ_n to μ.



(a) Points sampled i.i.d.. **(b)** Points sampled deterministically.

Figure 8.2: Illustration of the different notions of convergence of measures.

8.3.1 Preliminaries

We begin by giving some background on kernels, convergence in total variation, and Hölder's Inequality.

Definition 8.3.1. Given a bounded metric measure space (X, d, μ), its *associated kernel* $K\colon X \times X \to \mathbb{R}$ is

$$K(x,y) = \frac{1}{2}\left(-d(x,y)^2 + \int_X d(w,y)^2\mu(\mathrm{d}w) + \int_X d(x,z)^2\mu(\mathrm{d}z) - \int_{X\times X} d(w,z)^2\mu(\mathrm{d}w \times \mathrm{d}z)\right),$$

and its *associated linear operator* $T_K\colon L^2(X,\mu) \to L^2(X,\mu)$ is defined via

$$[T_K\phi](x) = \int_X K(x,y)\phi(y)\mu(\mathrm{d}y).$$

Definition 8.3.2 (Total-variation convergence of measures)**.** Let $(X, \mathcal{F})$ be a measurable space. The total variation distance between two (positive) measures μ and ν is then given by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_f \left\{\int_X f\, d\mu - \int_X f\, d\nu\right\}.$$

The supremum is taken over f ranging over the set of all measurable functions from X to $[-1, 1]$. In our definition of metric measure spaces, we consider Borel probability measure on the space X.

Indeed, convergence of measures in total-variation implies convergence of integrals against bounded measurable functions, and the convergence is uniform over all functions bounded by any fixed constant.

Theorem 8.3.3 (Hölder's Inequality)**.** *Let (S, Σ, μ) be a measure space and let $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then, for all measurable real- or complex-valued functions f and g on S,*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Lemma 8.3.4. *For a measure space X with finite measure ($\mu(X) < \infty$), $L^2(X)$ is contained in $L^1(X)$.*

Proof. Using the Schwarz' inequality (Hölder's inequality for $p, q = 2$), we have the following:

$$\int_X |f(x)|\mu(\mathrm{d}x) = \int_X 1 \cdot |f(x)|\mu(\mathrm{d}x) \leq \|1\|_2 \|f\|_2 < \infty,$$

since $\int_X 1\mu(\mathrm{d}x) = \mu(X) < \infty$. Thus, $L^2(X) \subseteq L^1(X)$. □

8.3.2 Convergence of MDS for Finite Measures

Let X be a bounded metric space, and let $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ for $n \geq 1$ be a sequence of averages of Dirac deltas with $\{x_1, \ldots, x_n\} \subseteq X$, and with μ_n converging to μ in total variation as $n \to \infty$. Each μ_n has finite support whereas μ may have infinite support. Let K_n and T_{K_n} be the kernel and operator associated to (X, d, μ_n), and let K and T_K be the kernel and operator associated to (X, d, μ). Let $\lambda_{k,n}$ denote the kth eigenvalue of T_{K_n} with associated eigenfunction $\phi_{k,n}$. Furthermore, let λ_k denote the kth eigenvalue of T_K with associated eigenfunction f_k. Following the proofs of [2, 16], we will show that the MDS embeddings of the metric measure space (X, d, μ_n) converge to the MDS embedding of (X, d, μ) in this deterministic setting.

By definition of K and K_n, they are both bounded since d is a bounded L^2-kernel with respect to the measures μ and μ_n for all n. It follows from the total variation convergence of $\mu_n \to \mu$ that $\|K_n - K\|_\infty \to 0$.

We use the result from [16, Theorem 3.1], where instead of studying finite spaces X_n, we study the possibly infinite space X equipped with a measure μ_n of finite support. Since μ_n has finite support, Theorem 8.2.3 still holds under the assumptions of this section.

Suppose for each k, the eigenfunction $\phi_{k,n}$ of T_{K_n} converges uniformly to some function $\phi_{k,\infty}$ as $n \to \infty$. In Proposition 8.3.6, we show that $\phi_{k,\infty}$ are the eigenfunctions of T_K.

Lemma 8.3.5. *$\phi_{k,\infty}$ is bounded.*

Proof. We first show that $\phi_{k,n}(x)$ is bounded. Indeed,

$$|\phi_{k,n}(x)| = \left| \frac{1}{\lambda_{k,n}} \int_X K_n(x,y)\phi_{k,n}(y)\mu_n(\mathrm{d}y) \right| \leq \frac{1}{|\lambda_{k,n}|} \int_X |K_n(x,y)|\,|\phi_{k,n}(y)|\,\mu_n(\mathrm{d}y)$$
$$\leq \frac{c}{|\lambda_{k,n}|} \int_X |\phi_{k,n}(y)|\,\mu_n(\mathrm{d}y).$$

The second inequality follows from the fact that K_n is bounded by some constant c. Furthermore, $\phi_{k,n}(x) \in L^2(X)$ and $\mu_n(X) = 1 < \infty$. It follows from Lemma 8.3.4 that $\phi_{k,n}(x) \in L^1(X)$, i.e. that $\int_X |\phi_{k,n}(y)|\,\mu_n(\mathrm{d}y) < \infty$. Furthermore, knowing that $\|\phi_{k,n} - \phi_{k,\infty}\|_\infty \to 0$ and $\ell^2(\lambda(T_{K_n}), \lambda(T_K)) \to 0$, we deduce that $\phi_{k,\infty}$ is bounded. □

Proposition 8.3.6. *Suppose $\mu_n = \frac{1}{n}\sum_{x \in X_n} \delta_x$ converges to μ in total variation. If the eigenfunctions $\phi_{k,n}$ of T_{K_n} converge uniformly to $\phi_{k,\infty}$ as $n \to \infty$, then their limit are the corresponding eigenfunctions of T_K.*

Proof. We have the following,

$$
\begin{aligned}
\phi_{k,n}(x) =& \frac{1}{\lambda_{k,n}} \int_X K_n(x,y)\phi_{k,n}(y)\mu_n(\mathrm{d}y) \\
=& \frac{1}{\lambda_k} \int_X K(x,y)\phi_{k,\infty}(y)\mu_n(\mathrm{d}y) \\
&+ \frac{\lambda_k - \lambda_{k,n}}{\lambda_{k,n}\lambda_k} \int_X K(x,y)\phi_{k,\infty}(y)\mu_n(\mathrm{d}y) \\
&+ \frac{1}{\lambda_{k,n}} \int_X \Big(K_n(x,y) - K(x,y)\Big)\phi_{k,\infty}(y)\mu_n(\mathrm{d}y) \\
&+ \frac{1}{\lambda_{k,n}} \int_X K_n(x,y)\Big(\phi_{k,n}(y) - \phi_{k,\infty}(y)\Big)\mu_n(\mathrm{d}y).
\end{aligned}
$$

By Lemma 8.3.5, $\phi_{k,\infty}$ is bounded. Therefore, we can insert $\frac{1}{\lambda_k}\int_X K(x,y)\phi_{k,\infty}(y)\mu(\mathrm{d}y)$ into the above aligned equations in order to obtain,

$$
\begin{aligned}
&\left|\phi_{k,n}(x) - \frac{1}{\lambda_k}\int_X K(x,y)\phi_{k,\infty}(y)\mu(\mathrm{d}y)\right| \\
\leq& \left|\frac{1}{\lambda_k}\int_X K(x,y)\phi_{k,\infty}(y)\mu_n(\mathrm{d}y) - \frac{1}{\lambda_k}\int_X K(x,y)\phi_{k,\infty}(y)\mu(\mathrm{d}y)\right| \\
&+ \left|\frac{\lambda_k - \lambda_{k,n}}{\lambda_{k,n}\lambda_k}\int_X K(x,y)\phi_{k,\infty}(y)\mu_n(\mathrm{d}y)\right| \\
&+ \left|\frac{1}{\lambda_{k,n}}\int_X \Big(K_n(x,y) - K(x,y)\Big)\phi_{k,\infty}(y)\mu_n(\mathrm{d}y)\right| \\
&+ \left|\frac{1}{\lambda_{k,n}}\int_X K_n(x,y)\Big(\phi_{k,n}(y) - \phi_{k,\infty}(y)\Big)\mu_n(\mathrm{d}y)\right| \\
:=& A_n + B_n + C_n + D_n.
\end{aligned}
$$

Since the $\lambda_{k,n}$ converge to λ_k, since the K_n converge to K, since the $\phi_{k,n}$ converge to $\phi_{k,\infty}$, and since $\phi_{k,\infty}$, K, and K_n are bounded, it follows that the B_n, C_n, and D_n converge to 0 as $n \to \infty$.

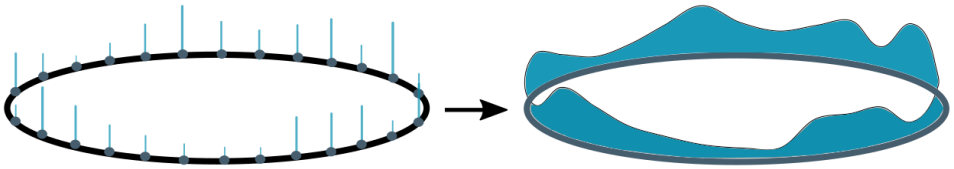
Since $\mu_n \to \mu$, we also have that $A_n \to 0$ as $n \to \infty$ Therefore

$$\phi_{k,n}(x) \to \frac{1}{\lambda_k} \int_X K(x,y)\phi_{k,\infty}(y)\mu(\mathrm{d}y) = \frac{1}{\lambda_k}[T_k\phi_{k,\infty}](x)$$

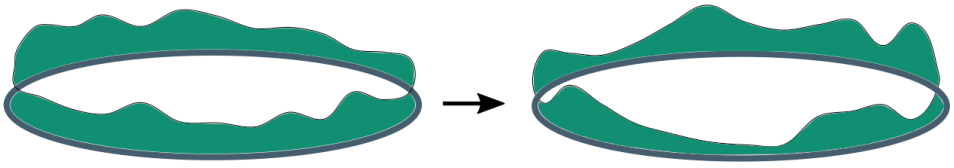
for all $x \in X$. Since we also have $\phi_{k,n}(x) \to \phi_{k,\infty}(x)$, it follows that $\lambda_k\phi_{k,\infty}(x) = T_k\phi_{k,\infty}$. Therefore $\phi_{k,\infty}$ is an eigenfunction of T_K with an eigenvalue λ_k. □

8.4 Convergence of MDS for Arbitrary Measures

We now generalize the setting of Section 8.3 to allow for arbitrary measures, as opposed to finite sums of Dirac delta measures as illustrated in Figure 8.3a. Suppose X is a bounded metric space, and μ_n is an arbitrary sequence of probability measures on X_n for all $n \in \mathbb{N}$, such that μ_n converges to μ in total variation as $n \to \infty$. For example, the support of each μ_n is now allowed to be infinite as illustrated in Figure 8.3b. We will give some first results towards showing that the MDS embeddings of (X, d, μ_n) converge to the MDS embedding of (X, d, μ).



(a) Convergence of arbitrary measures with finite support.



(b) Convergence of arbitrary measures with infinite support.

Figure 8.3: Illustration of convergence (in total variation) of arbitrary measures.

The bounded metric measure space (X, d, μ) is equipped with a kernel $K\colon X \times X \to \mathbb{R}$ and linear operator $T_K\colon L^2(X,\mu) \to L^2(X,\mu)$, as defined as in Definition 8.3.1. For (X, d, μ_n), we denote the analogous kernel by $K_n\colon X \times X \to \mathbb{R}$ and its linear operator by $T_{K_n}\colon L^2(X,\mu_n) \to L^2(X,\mu_n)$. Let $\lambda_{k,n}$ denote the kth eigenvalue of T_{K_n} with associated eigenfunction $\phi_{k,n}$. Furthermore, let λ_k denote the kth eigenvalue of T_K with associated eigenfunction ϕ_k.

Proposition 8.4.1. *Suppose μ_n converges to μ in total variation. If the eigenvalues $\lambda_{k,n}$ of T_{K_n} converge to λ_k, and if their corresponding eigenfunctions $\phi_{k,n}$ of T_{K_n} converge uniformly to $\phi_{k,\infty}$ as $n \to \infty$, then the $\phi_{k,\infty}$ are eigenfunctions of T_K with eigenvalue λ_k.*

Proof. The same proof of Proposition 8.3.6 holds. Indeed, so long as we know that the eigenvalues $\lambda_{k,n}$ of T_{K_n} converge to λ_k, then nowhere else in the proof of Proposition 8.3.6 does it matter whether μ_n is an average of Dirac delta masses or instead an arbitrary probability measure. □

We conjecture that the hypothesis in Proposition 8.4.1 about the convergence of eigenvalues is unnecessary.

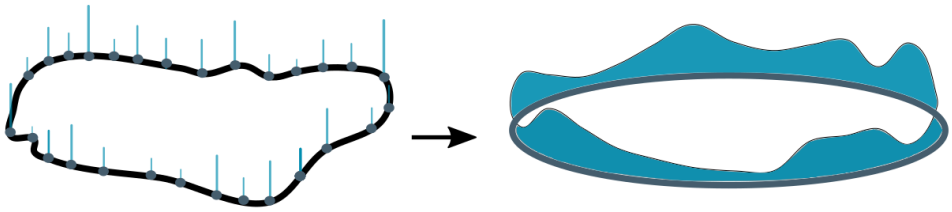
Conjecture 8.4.2. Suppose we have the convergence of measures $\mu_n \to \mu$ in total variation. The ordered spectrum of T_{K_n} converges to the ordered spectrum of T_K as $n \to \infty$ with respect to the ℓ^2–distance,

$$\ell^2(\lambda(T_{K_n}), \lambda(T_K)) \to 0.$$

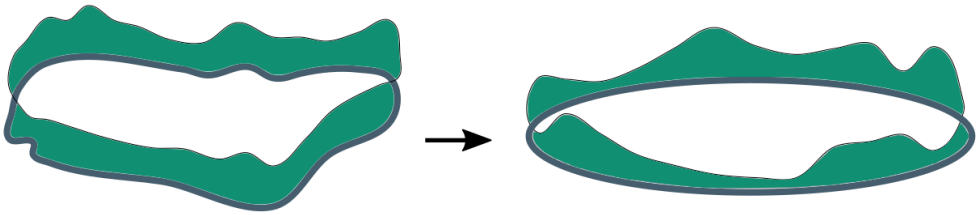
Remark 8.4.3. We remark that some ideas from the proof of [16, Theorem 3.1] may be useful here. One change is that the inner products considered in equation (3.4) of the proof of [16, Theorem 3.1] may need to be changed to inner products with respect to the measure μ_n.

8.5 Convergence of MDS with Respect to Gromov–Wasserstein Distance

We now consider the more general setting in which (X_n, d_n, μ_n) is an arbitrary sequence of metric measure spaces, converging to (X, d, μ) in the Gromov–Wasserstein distance as illustrated in Figure 8.4a for the finite case and Figure 8.4b for the infinite case. We remark that X_n need to no longer equal X, nor even be a subset of X. Indeed, the metric d_n on X_n is allowed to be different from the metric d on X. Sections 8.2 and 8.3 would be the particular case (depending on your perspective) when either X_n is a finite subset of X and d_n is the restriction of d, or equivalently when (X_n, d_n) is equal to (X, d) but μ_n is a finite average of Dirac delta masses. Section 8.4 is the particular case when $(X_n, d_n) = (X, d)$ for all n, and the measures μ_n are converging to μ. We now want to consider the case where metric d_n need no longer be equal to d.



(a) Convergence of mm-spaces equipped with measures of finite support.



(b) Convergence of mm-spaces equipped with measures of infinite support.

Figure 8.4: Illustration of Gromov–Wasserstein convergence of arbitrary metric measure spaces (mm-spaces).

Conjecture 8.5.1. Let (X_n, d_n, μ_n) for $n \in \mathbb{N}$ be a sequence of metric measure spaces that converges to (X, d, μ) in the Gromov–Wasserstein distance. Then the MDS embeddings converge.

Question 8.5.2. Are there other notions of convergence of a sequence of arbitrary (possibly infinite) metric measure spaces (X_n, d_n, μ_n) to a limiting metric measure space (X, d, μ) that would imply that the MDS embeddings converge in some sense? We remark that one might naturally try to break this into two steps: first analyze which notions of convergence $(X_n, d_n, \mu_n) \to (X, d, \mu)$ imply that the operators $T_{K_n} \to T_K$ converge, and then analyze which notions of convergence on the operators $T_{K_n} \to T_K$ imply that their eigendecompositions and MDS embeddings converge.

Chapter 9

Conclusion

MDS is concerned with problem of mapping the objects $x_1, \ldots, x_n$ to a configuration (or embedding) of points $f(x_1), \ldots, f(x_n)$ in $\mathbb{R}^m$ in such a way that the given dissimilarities d_{ij} are well-approximated by the Euclidean distances between $f(x_i)$ and $f(x_j)$. We study a notion of MDS on infinite metric measure spaces, which can be simply thought of as spaces of (possibly infinitely many) points equipped with some probability measure. We explain how MDS generalizes to infinite metric measure spaces. Furthermore, we describe in a self-contained fashion an infinite analogue to the classical MDS algorithm. Indeed, classical multidimensional scaling can be described either as a Strain-minimization problem, or as a linear algebra algorithm involving eigenvalues and eigenvectors. We describe how to generalize both of these formulations to infinite metric measure spaces. We show that this infinite analogue minimizes a Strain function similar to the Strain function of classical MDS. This theorem generalizes [3, Theorem 14.4.2], or equivalently [37, Theorem 2], to the infinite case. Our proof is organized analogously to the argument in [37, Theorem 2].

As a motivating example for convergence of MDS, we consider the MDS embeddings of the circle equipped with the (non-Euclidean) geodesic metric. By using the known eigendecomposition of circulant matrices, we identify the MDS embeddings of evenly-spaced points from the geodesic circle into $\mathbb{R}^m$, for all m. Indeed, the MDS embeddings of the geodesic circle are closely related to [39], which was written prior to the invention of MDS.

Lastly, we address convergence questions for MDS. Indeed, convergence is well-understood when each metric space has the same finite number of points [31], but we are also interested in convergence when the number of points varies and is possibly infinite. We survey Sibson's perturbation analysis [31] for MDS on a fixed number of n points. We survey results of [2, 16] on

the convergence of MDS when n points $\{x_1, \ldots, x_n\}$ are sampled from a metric space according to a probability measure μ, in the limit as $n \to \infty$. We reprove these results under the (simpler) deterministic setting when points are not randomly chosen, and instead we assume that the corresponding finite measures $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ (determined by n points) converge to μ. This allows us, to consider the more general setting where we have convergence of *arbitrary* probability measures $\mu_n \to \mu$. However, several questions remain open. In particular, we would like to have a better understanding of the convergence of MDS under the most unrestrictive assumptions of a sequence of arbitrary (possibly infinite) metric measure spaces converging to a fixed metric measure space, perhaps in the Gromov–Wasserstein distance (that allows for distortion of both the metric and the measure simultaneously); see Conjecture 8.5.1 and Question 8.5.2.

Despite all of the work that has been done on MDS by a wide variety of authors, many interesting questions remain open (at least to us). For example, consider the MDS embeddings of the n-sphere for $n \geq 2$.

Question 9.0.1. What are the MDS embeddings of the n-sphere S^n, equipped with the geodesic metric, into Euclidean space $\mathbb{R}^m$?

To our knowledge, the MDS embeddings of S^n into $\mathbb{R}^m$ are not understood for all positive integers m except in the case of the circle, when $n = 1$. The above question is also interesting, even in the case of the circle, when the n-sphere is not equipped with the uniform measure. As a specific case, what is the MDS embedding of S^1 into $\mathbb{R}^m$ when the measure is not uniform on all of S^1, but instead (for example) uniform with mass $\frac{2}{3}$ on the northern hemisphere, and uniform with mass $\frac{1}{3}$ on the southern hemisphere?

We note the work of Blumstein and Kvinge [5], where a finite group representation theoretic perspective on MDS is employed. Adapting these techniques to the analytical setting of compact Lie groups may prove fruitful for the case of infinite MDS on higher dimensional spheres.

We also note the work [6], where the theory of an MDS embedding into pseudo Euclidean space is developed. In this setting, both positive and negative eigenvalues are used to create an embedding. In the example of embedding S^1, positive and negative eigenvalues occur in a one-to-one fashion. We wonder about the significance of the full spectrum of eigenvalues for the higher dimensional spheres.

Bibliography

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[2] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, Jean-François Paiement, Pascal Vincent, and Marie Ouimet. Learning eigenfunctions links spectral embedding and kernel PCA. *Neural computation*, 16(10):2197–2219, 2004.

[3] JM Bibby, JT Kent, and KV Mardia. Multivariate analysis, 1979.

[4] Leonard Mascot Blumenthal. *Theory and applications of distance geometry*. Chelsea New York, 1970.

[5] Mark Blumstein and Henry Kvinge. Letting symmetry guide visualization: Multidimensional scaling on groups. *arXiv preprint arXiv:1812.03362*, 2018.

[6] Mark Blumstein and Louis Scharf. Pseudo Riemannian multidimensional scaling. 2019.

[7] Eugène Bogomolny, Oriol Bohigas, and Charles Schmit. Spectral properties of distance matrices. *Journal of Physics A: Mathematical and General*, 36(12):3595, 2003.

[8] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[9] Andreas Buja, Deborah F Swayne, Michael L Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, 2008.

[10] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. CRC press, 2000.

[11] Jon Dattorro. *Convex optimization & Euclidean distance geometry*. Lulu.com, 2010.

[12] Jan de Leeuw and Willem Heiser. 13 theory of multidimensional scaling. *Handbook of statistics*, 2:285–316, 1982.

[13] Persi Diaconis, Sharad Goel, Susan Holmes, et al. Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, 2(3):777–807, 2008.

[14] PJ Groenen and Ingwer Borg. Past, present, and future of multidimensional scaling. *Visualization and verbalization of data*, pages 95–117, 2014.

[15] Ram P Kanwal. *Linear integral equations*. Springer Science & Business Media, 2013.

[16] Vladimir Koltchinskii, Evarist Giné, et al. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

[17] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[18] Thomas Kühn. Eigenvalues of integral operators generated by positive definite hölder continuous kernels on metric compacta. In *Indagationes Mathematicae (Proceedings)*, volume 90, pages 51–61. Elsevier, 1987.

[19] Henry Kvinge, Elin Farnell, Michael Kirby, and Chris Peterson. A GPU-oriented algorithm design for secant-based dimensionality reduction. In *2018 17th International Symposium on Parallel and Distributed Computing (ISPDC)*, pages 69–76. IEEE, 2018.

[20] Kanti V Mardia. Some properties of classical multi-dimensional scaling. *Communications in Statistics-Theory and Methods*, 7(13):1233–1241, 1978.

[21] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[22] Facundo Mémoli. The Gromov–Wasserstein distance: A brief overview. *Axioms*, 3(3):335–341, 2014.

[23] Elzbieta Pekalska, Pavel Paclik, and Robert PW Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of machine learning research*, 2(Dec):175–211, 2001.

[24] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[25] John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, 100(5):401–409, 1969.

[26] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[27] Matthias Scholz, Fatma Kaplan, Charles L Guy, Joachim Kopka, and Joachim Selbig. Non-linear pca: a missing data approach. *Bioinformatics*, 21(20):3887–3895, 2005.

[28] Roger N Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125–140, 1962.

[29] Roger N Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27(3):219–246, 1962.

[30] Robin Sibson. Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 234–238, 1978.

[31] Robin Sibson. Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 217–229, 1979.

[32] Robin Sibson, Adrian Bowyer, and Clive Osmond. Studies in the robustness of multidimensional scaling: Euclidean models and simulation studies. *Journal of Statistical Computation and Simulation*, 13(3-4):273–296, 1981.

[33] Frank Smithies. *Integral Equations*. Cambridge University Press, 1970.

[34] Karl-Theodor Sturm. On the geometry of metric measure spaces. *Acta mathematica*, 196(1):65–131, 2006.

[35] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[36] Andrew Timm. *An investigation of Multidimensional Scaling with an emphasis on the development of an R based Graphical User Interface for performing Multidimensional Scaling procedures*. PhD thesis, University of Cape Town, 2012.

[37] Michael Trosset. Computing distances between convex sets and subsets of the positive semidefinite matrices. Technical report, 1997.

[38] Michael W Trosset. A new formulation of the nonmetric strain problem in multidimensional scaling. *Journal of Classification*, 15(1):15–35, 1998.

[39] John Von Neumann and Isaac Jacob Schoenberg. Fourier integrals and metric geometry. *Transactions of the American Mathematical Society*, 50(2):226–251, 1941.

[40] Wikipedia contributors. Functional principal component analysis Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Functional_principal_component_analysis, 2018. [Online; accessed 5-February-2019].

[41] Christopher KI Williams. On a connection between kernel PCA and metric multidimensional scaling. In *Advances in neural information processing systems*, pages 675–681, 2001.

[42] WA Wilson. On certain types of continuous transformations of metric spaces. *American Journal of Mathematics*, 57(1):62–68, 1935.