Comparing test statistics of two-sample permutation tests based on covariance of functional data

Hackmann, Toby

Universiteit
Leiden
The Netherlands

Comparing test statistics of two-sample permutation tests based on covariance of functional data

Toby Hackmann

Thesis advisor: Dr. V. Masarotto, Leiden University

Defended on August 22nd, 2023

MASTER THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

With modern measurement techniques, the prominence of data that can be best described as *functional data* is increasing. Statistical techniques for comparing the means of two samples of functional data are well studied, but less is known about tests for comparisons of covariance. We suggest a new method to analyze covariances, making use of an extension of the Hilbert-Schmidt space of operators with trace-class elements and defining new metrics on that extended space. Then, we compare these new methods with other known methods in both an analysis of a data set, and a simulation study. Our results show that, in a two-sample permutation test of the phoneme data set from the `fda` package in `R`, the Riemannian (geodesic path length) and log-Euclidean test statistics perform by far the best, with a power of 1.000 for all three two-sample comparisons. By contrast, the Frobenius and Procrustes test statistics had powers between 0.5 and 1, depending on the sample. A simulation study shows that the power of a two-sample permutation test can differ vastly depending on the type difference between the covariances of the two samples. For some covariance functions, the power of the Frobenius test statistic was close to 1, while that of the Riemannian and log-Euclidean was much lower. But for other covariance functions, the reverse was true. This means that a test statistic must be chosen very carefully by researchers.

Summary

Data used in classical statistical settings can be seen as individual measurements. For example, measurements of two groups of subjects that have been exposed to a different treatment. Functional data, in contrast, is data of a continuous nature, such as measurements over time. Examples of this type of data are weather measurements, such as temperature curves, which are often continuous. With this type of data, we can also compare two groups subjected to different treatments.

If we want to see if two sets of functions are different, we can do that by comparing their average, also called the mean. Most research has been done on comparing the means of functions. However, the difference between two sets of functions could be in their spread and *wiggliness* around its mean function, instead of the locations of the means. We call this spread and wiggliness the *covariance* of a sample.

There are quite a few known methods that can be used to do such an analysis. Most rely on the practical fact that we can only measure functions at a finite number of points, while it is theoretically possible to do so at infinitely many points. A consequence of these methods is that, when you use them to estimate the covariance, the estimated covariance no longer has the properties that you expect from a covariance. We attempt to solve this by suggesting a new mathematical method that does respect the properties that arise from potentially infinite dimensional data.

Then, we compare the estimators from this new method to established methods. We do this in both an analysis of an actual data set and on data that we simulate ourselves from a chosen covariance. For such a comparison, we choose to use a permutation test. Permutation tests for two samples compare a value, called the test statistic, that is calculated based on the difference between the actual two samples. This is then compared to the test statistics for many random permutations of the data put into the two groups. The comparison will be based on the so-called 'power' of a test. The power of a test is the probability that a statistical test gives you a statistically significant result, if your hypothesis is true.

From this comparison of powers, we have found that using the new methods to calculate the estimated covariance of a sample does not work with our computational methods. The results from that are the same as we would expect from randomly choosing an estimate. With some of the other methods, we have found some very interesting results. Using one of the metrics, that calculates the shortest path over a surface (like over the surface of the earth instead of through it), we have found a power of 1, which would mean that we are essentially guaranteed to get a statistically significant results if we would test for those differences. The other, more

standard distance functions behaved as expected. Their results showed us that the covariances were easier to tell apart if they were more different.

Our simulation study confirmed these findings. The most important result here is, that some of the test statistics work much better than others in some cases and much worse in others. This means that there is no 'best' test to use in all situations. Future directions for research could be into the scenarios where each of these test statistics work best. That way, researches can be pointed in the right direction when using these tests. Especially in the simulations with vector parameters, that made the covariance function asymmetric along the diagonal, there are big differences in the power of the permutation tests. With some parameters, the Riemannian and log-Euclidean test statistics performed extremely well and the Frobenius and Procrustes metrics rather poorly. With other parameters, these results were completely the other way around.

The structure of the thesis is as follows. We start by giving a general introduction to the topic of functional data and our current knowledge of its analysis. From there, we move on to an introduction to the mathematical background that is at the foundation of functional data analysis. In the third chapter, we continue with the introduction of a new metric that could be used to analyze functional data. After that, we put the mathematics to the side and look into a proof of concept data analysis to see how the different metrics perform in an analysis of a data set. Following the data analysis, we want to increase our understanding of the results by looking at it in more depth in a simulation study in chapter five. The thesis will then be concluded with a discussion and conclusion.

Contents

Chapter 1

Introduction

Functional Data Analysis (fda) is a branch of statistics that deals with the analysis of data that are continuous in nature, such as time series or functional observations. The goal of fda is to model, analyze, and make inferences about functional data, which are often represented as curves or surfaces. Often, these functions are something that is continuous over time, but it can also be continuous over some else such as space or probability. Examples of functional data over time are time series, such as stock market data for economics or growth data in children. An example of functional data with no time component is the dataset that we will be using, containing speech patterns of several phonemes. Phoneme data is a function over frequency, where the full function is relevant for analysis. As can be seen in figure 1, any two phonemes are very close together at some point of the function, and they can only be clearly differentiated over the full range.

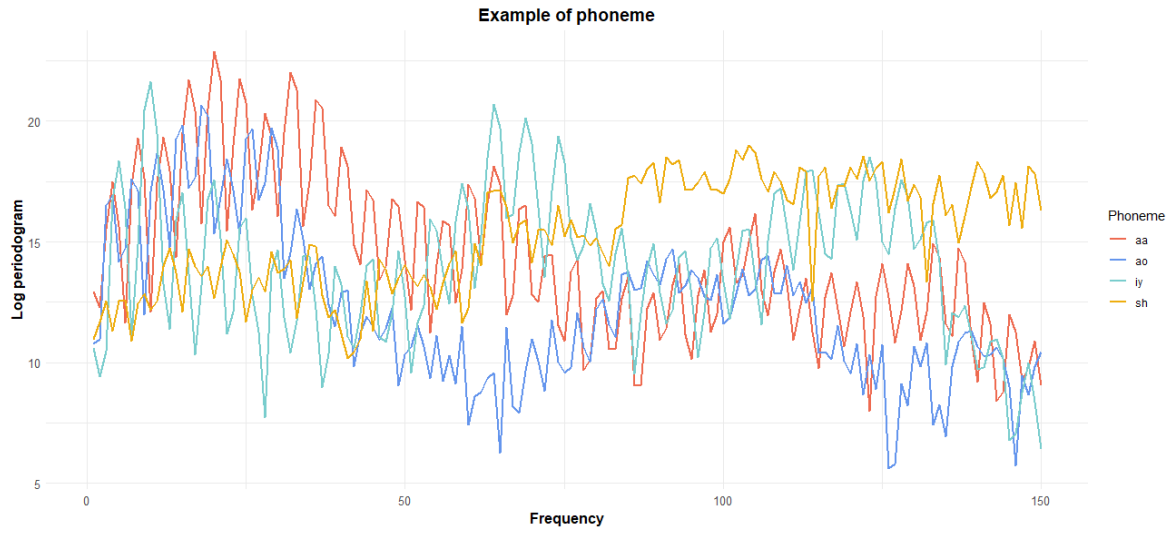


Figure 1: Examples of the four phoneme that we will be comparing later on in this thesis. This is an example where the function values, in this case the log periodogram, are not a function of time or space, but of frequency. A periodogram is the measure of spectral density of a signal, in this case the pronunciation of a phoneme. These graphs therefore show which frequencies occur more in the different phoneme.

Another area in which functional data has become increasingly relevant is in biomedical

sciences. In particular, the emergence of live cell imaging techniques (Cole, 2014), contributes to the rapid increasing dimensionality of datasets. With this increasing dimensionality, functional data analysis becomes an increasingly important tool for analysing this data. An example of the use of techniques derived from fda can be seen in Vu et al. (2022), where the Mark Connection Function (mcf) is used as a measurement of how clustered cells of a certain type are. The two-dimensional image with cell location and labels is transformed into a function of randomness of the cell pattern over resolution, as seen in figure 2.

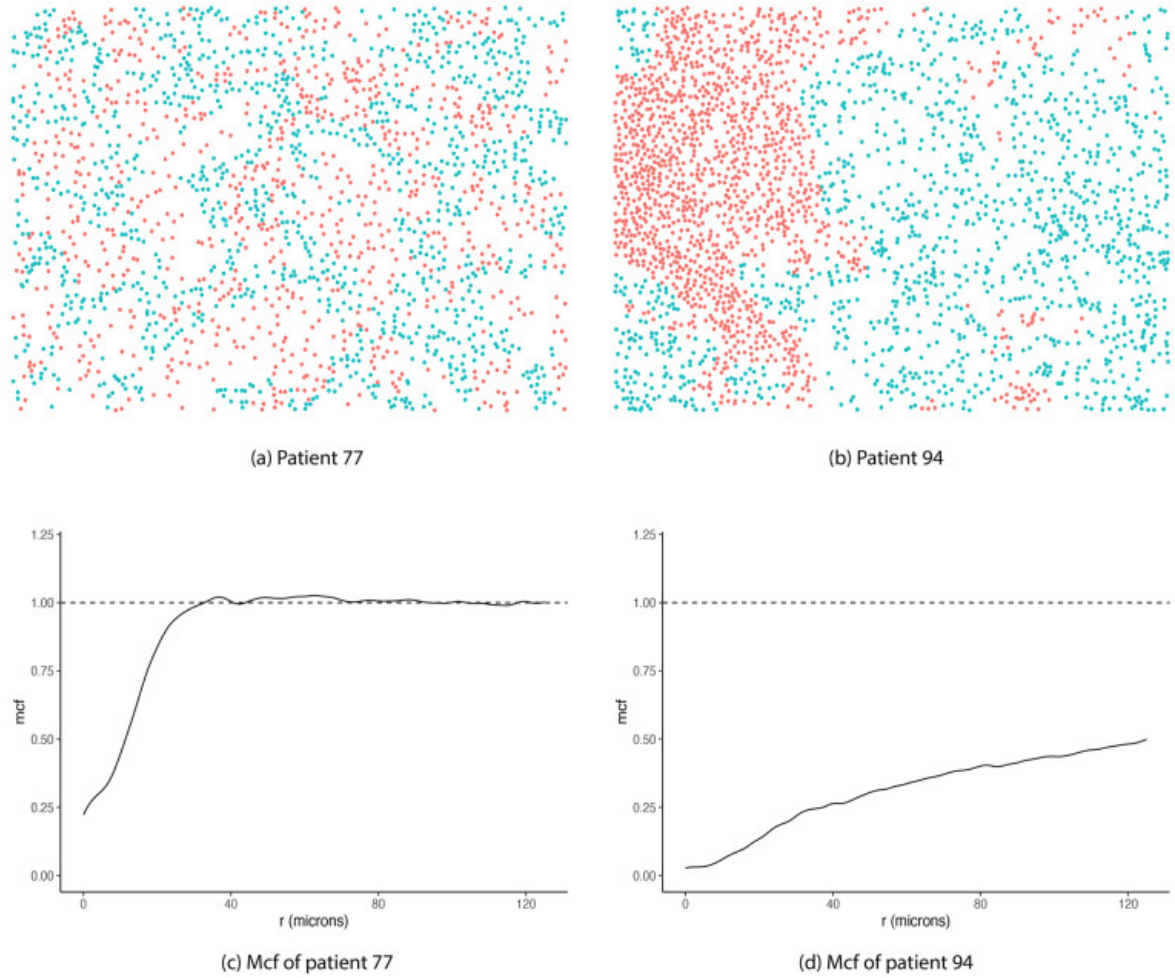


Figure 2: Examples of mcf functions. On the top row, we have the two-dimensional data of the cell locations and their types in color. In teal, we have the tumor cells and in red the healthy cells. The bottom two images show the mcf functions for the respective spatial data on top. On the left, we can clearly see that the cells are not very clustered and the mcf line quickly approaches 1, which represents a random distribution of cells. For the clustered cells on the right, the mcf function stays much lower with increasing resolution. (Vu et al., 2022)

These functional summaries of a key data characteristic, such as the mcfs can be much more easily compared than the underlying data. Comparisons of the means of samples of functional data have been studied extensively and the asymptotic distributions of their test statistics are known in many cases, such as those explored by Jiménez Gamero and Franco Pereira (2021). However, only comparing the means of functional data sample ignores the vast treasure trove of information that is contained within the covariance of functions. Having good methods to compare the covariances of functional data lead to some great insights. An example of this can

be found in Panaretos et al. (2010), where they applied it to data on DNA minicircles.

The data for a second order (covariance) comparison is much harder to use. Covariances in general, are infinite dimensional operators on functions. In practice, with our finite dimensional datasets, they are $n \times n$ matrices, where n is the number of points where we know the function values. So the covariance of a function already has massively more data entries than the function itself. Computational actions are also much more resource intensive on matrices. A basic matrix multiplication algorithm of a square $n \times n$ matrix is of order $\mathcal{O}(n^3)$, meaning that by doubling the number of measurements of your function, the time it takes for a single covariance matrix multiplication is increased eight-fold. Faster algorithms do exist that reduce the order to $\mathcal{O}\left(n^{\log_2(7)}\right)$ (Strassen, 1969), or even further, as shown by Coppersmith and Winograd (1990). Clearly, this can lead to big computational issues as the dimensionality of functional data increases.

Compounding these issues with matrix multiplication is the type of statistical tests that are used with second order analysis. Not much is known about the asymptotic distributions of test statistics for covariance comparison. This means that we cannot simply analyze the test statistic by comparing it to a known distribution. In many cases, this means that permutation tests are required to get the results, and these are also computationally intensive.

Still, with modern systems, this is generally doable for smaller amounts of lower dimensional data. When the dimensionality of the functions increases, this does not only cause computational issues, but also issues on the mathematical side of the analysis. There are mostly caused by the inherent infinite dimensionality of functional data.

1.1 Analysis of functional data

Intrinsically, functional data is infinite dimensional. The data can generally be represented as a function $X : T \to \mathbb{R}$, where T is an index set and $\mathbb{R}$ is tbe set of real numbers. For simplicity, we only consider functions that take values in $\mathbb{R}$, since that covers most of the practical applications. In general, this index set is a closed subset of a real $\mathbb{R}$ space, since these processes usually have specified beginnings and ends. Since the set T is closed (there is a minimum and maximum value of T), we can scale this down to the set $[0, 1]$ without losing any information. In practice, functional data is often reduced to finite dimensionality for practical purposes. When data needs to be saved and processed, only a finite number of data can be stored. Instead of storing the full function $\{X(t) \mid t \in [0, 1]\}$, we store a selected number n of points $\left\{X\left(\frac{k}{n}\right) \mid k \in \{0, 1, \ldots, n\}\right\}$. While this reduces the dimensionality from infinity down to some finite number. However, even this finite number of dimensions is often still very large and poses the problems with computation mentioned before.

To overcome these issues with high dimensionality, techniques are often used to reduce the dimensionality of the data. By finding the most important features or patterns in the data, we can keep those and discard the rest of the data. This process is often referred to as dimension reduction. Some common techniques used in fda for dimension reduction include functional principal component analysis (Ramsay & Silverman, 2005), functional linear discriminant anal-

ysis (James & Titterington, 2005), and functional canonical correlation analysis (Marquez et al., 2008). These techniques enable researchers to identify the most important features in the functional data, and make inferences about the underlying processes that generated the data.

Besides these inferential methods, machine learning techniques have also gained traction for fda. While many machine learning techniques are well suited to the task, recurrent neural networks in particular are very good for sequential data such as time series (Brown, 2020). For spatial functions, such as handwriting, recurrent neural networks have also proven their worth, as presented by Graves et al. (2013).

1.2 Metrics for covariance

All the mentioned techniques have pushed the field of functional data analysis forward, giving us more insight into the underlying processes. Many are still restricted to the practical situation with finite dimensional data. With the current amount of data that we can gather, these methods are still good. However, as technology improves, so does the resolution of functional data that we can gather and it is not guaranteed that methods that rely on finite dimensionality carry over well at increasing scale. Making the jump from a finite dimensional framework to an infinite dimensional one is not straightforward, and not all methods carry over nicely.

One of the most important fundamentals for all statistics is the idea of a distance function. Something as simple as taking the mean of a sample, is finding the point that minimizes the average distance to each of the points in the sample. We understand how this works in a finite dimensional setting, but it doesn't translate easily. How can we measure the distance between two points in an infinite dimensional space?

A typical question for functional data can be to distinguish between multiple populations of functions. Examples of this may be distinguishing between two sets of signatures in handwriting analysis, or two different vowels in analysis of phonemes. This can be expressed as finding differences between K groups, where these are represented as functions X_k for each $k \in K$. Each group has a mean function $\mu_k \in \mathbb{H}$ in a Hilbert space of functions and a covariance operator $\Sigma_k : \mathbb{H} \times \mathbb{H} \to \mathbb{H}$, between Hilbert spaces. For those unfamiliar, the definition of a Hilbert space will be discussed in chapter 2.1.3.

One way to differentiate between groups of functions is to compare their mean functions, but this is not always where differences are found. It is possible that the means of groups are very similar, but the differences between the groups are in the second order, also known as covariance. Differences in covariance express as different oscillations of the functions around their mean. While there are many good methods of finding the distance between two mean functions, it is much harder to define a distance function on covariance operators that respects the unique geometry of the space they exist in.

To get to a definition of distance between two covariances, we need to go into the mathematics of operators. We will then build on the work of Lawson and Lim (2013b) and propose a similar

method that respects all the properties inherent to an operator in an infinite dimensional space. Based on this new method, we will then propose metrics that can be used to compare two different samples of functional data based on their covariances.

1.3 Power of covariance tests

It is great that it is theoretically possible to distinguish two samples of data based on their covariance. To see the effectiveness of covariance tests, we will test the power of a two-sample permutation test. This will be done on both the `phoneme` (Hastie et al., 1995) data set contained in the `fda` package in `R` and on simulated data.

We use the phoneme data, because other studies have also used that as an example data set, such as Masarotto et al. (2022), Hlávka et al. (2022) and Kashlak et al. (2018), while Pigoli et al. (2014) used a different data set of spoken languages. Kashlak et al. (2018) also performed k-sample power tests on the phoneme data, while we will simply use a two-sample test of differences.

As mentioned, we will mostly focus on the power of the permutation tests. The power of a test is an important metric for analyzing the quality of a test. If a statistical test has a low power, then it is basically a gamble if a study gets a statistically significant result, even if the alternative hypothesis is true. When designing a study, it is therefore important to choose a test with a power that is as high as possible. Knowing what power you can expect given a certain sample size is also important when designing studies for the lowest possible number of participants.

Chapter 2

Mathematical Background

As with most statistics, it is useful to start from probability theory. We adopt the framework, in which objects that are studied in fda are stochastic processes. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a non-empty index set T, a stochastic process can be defined as

$$X = \{X(\omega, t) \mid t \in T,\ \omega \in \Omega\}, \tag{2.1}$$

where $X(\ .\ , t)$ is a $\mathcal{F}$-measurable function. The ω refers to which of the possible stochastic processes in the sample space X is, but this argument is usually dropped. The index set T is usually a bounded interval of $\mathbb{R}$ in the context of fda, so we will assume this. When it is such a bounded interval, T can be reduced to the set $[0, 1]$ without loss of information. This makes the set much easier to work with. This is a mathematical description of these stochastic processes. Where can we find these processes, in which spaces? How can we define a distance between functions?

2.1 Function spaces

As mentioned earlier, the concept of distances is fundamental to statistics. To understand what is required of a distance function, or metric, for covariance operators, we will briefly overview metric spaces, normed vector spaces and Banach spaces. From there we will introduce Hilbert spaces, where the functions that are studied can be found. The next step is the introduction of Hilbert-Schmidt spaces, of which a subspace is the home of covariance operators.

Do note that we assume basic knowledge of linear algebra. For those that are looking for a good introduction to linear algebra, Strang (2009) is suggested. A brief overview for those that are familiar with the material, but want a refresher related to fda, Cai and Hsing (2015) is a good choice.

2.1.1 Metrics and normed vector spaces

The concept of distance has already been mentioned several times. Mathematically, distance is often measured using *metrics*. This is a standardization of the concept with the following definition:

Definition 2.1. *A **metric** on a set $\mathbb{M}$ is a function $d : \mathbb{M} \times \mathbb{M} \to \mathbb{R}$ that satisfies:*

1. *$d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$,*
2. *$d(x, y) = d(y, x)$, and*
3. *$d(x, y) \leq d(x, z) + d(y, z)$*

for $x,\ y,\ z \in \mathbb{M}$. The pair $(\mathbb{M}, d)$ is called a metric space.

Some important properties are immediately clear. Metrics are a distance function and as such, they can not be negative. They are also only 0 when both elements are the same. Metrics are symmetric, which makes perfect sense in most intuitive ideas of distance. This property could be an issue in some situations, such as considering the driving distance when considering one-way streets. The third part of the definition is more commonly known as the triangle inequality. The direct route between two points is always shorter than a detour through a third point z.

Metrics are very closely related to the concept of a norm. Instead of measuring distance, a norm measures the size of a vector in a vector space. It can be defined as follows:

Definition 2.2. *Given a vector space $\mathbb{V}$, a **norm** is a function $\|.\| : \mathbb{V} \to \mathbb{R}$ that satisfies:*

1. *$\|v\| \geq 0$, and $\|v\| = 0$ if and only if $v = 0$,*
2. *$\|av\| = |a|\|v\|$, and*
3. *$\|v + w\| \leq \|v\| + \|w\|$,*

for $v, w \in \mathbb{V}$ and $a \in \mathbb{R}$. The pair $(\mathbb{V}, \|.\|)$ is called a normed vector space.

From the definition, it is clear that the concept of norm is very similar to that of the metric. It even has a similar inequality as the metric as the final part of its definition, also referred to as triangle inequality. They can directly be linked through the following theorem:

Theorem 2.3. *Given a vector space $\mathbb{V}$ with norm $\|.\|$, $d(v, w) := \|v - w\|$ is a metric for $v, w \in \mathbb{V}$.*

This follows directly from checking the properties of the norm and metric. Norms and metrics are often quite intuitive in $\mathbb{R}^n$, where the Euclidean norm and metric provide an easy and intuitive understanding. How these concepts translate to function spaces can be more difficult to grasp, but the mathematics make it easy to check if this translation is valid.

Consider the function space $C[0, 1]$ of continuous functions on the interval $[0, 1]$. One norm that can be defined on this space is the 1-norm, given by

$$\|f\|_1 = \int_0^1 |f(x)|dx, \tag{2.2}$$

for f in $C[0, 1]$. Using theorem 2.3, we find that the metric that is defined by this norm is the integration metric

$$d(f, g) = \int_0^1 |f(x) - g(x)|dx, \tag{2.3}$$

for $f, g \in C[0, 1]$. This allows us to define a distance between two functions. Note that since we take the absolute value, the integral is always greater than or equal to 0. Because of the

absolute value, the integral is an increasing function over the interval and if the functions are ever different, the metric is positive. This means it is only 0, precisely when $f = g$. The other two conditions can be verified almost as easily.

2.1.2 Banach spaces

Now that clear definitions of both metrics and norms are established, we can use these elements in further constructions. One such use, that is quite important for convergence, is the construction of sequences of elements that get closer together. More formally,

Definition 2.4. *A sequence of elements $\{x_n\}$ of a metric space $(\mathbb{M}, d)$ is called a **Cauchy** sequence if for all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$, such that for all $n, m \geq N$ it holds that $d(x_n, x_m) < \varepsilon$.*

Of course, it would be very preferable if these Cauchy sequences converged to some element in our metric space. If the Cauchy sequence were one of functions, slowly moving towards the population mean, it is important that the element it converges to is also a valid function. Unfortunately, this is not true for every metric space.

Definition 2.5. *A metric space $(\mathbb{M}, d)$, where every Cauchy sequence converges, is called **complete**.*

Note that completeness of a metric space depends on both the set $\mathbb{M}$ and the metric d. A metric space can be complete with one metric, but not with another. The example of the $\|.\|_1$ on $C[0, 1]$ does constitute a complete normed vector space. Complete normed vector spaces have been given a specific name.

Definition 2.6. *A normed vector space that is complete under the metric associated with its norm is called a **Banach** space.*

As mentioned, the normed vector space $(C[0, 1], \|.\|_1)$ is a Banach space, which is a requirement of the space for functions for fda. There are many Banach spaces with functions, but the most relevant are usually the following spaces:

Definition 2.7. *Let $(S, \mathscr{A}, \mu)$ be a measure space and let $1 \leq p < \infty$. Define $\mathbb{L}^p(S, \mathscr{A}, \mu)$ as the collection of measurable functions f on S that satisfy*

$$\|f\|_p = \left(\int_S |f|^p \, d\mu \right)^{\frac{1}{p}} < \infty. \tag{2.4}$$

For $p = \infty$, define $\mathbb{L}^p(S, \mathscr{A}, \mu)$ as the collection of functions that are finite a.e., with

$$\lim_{p \to \infty} \|f\|_p = \|f\|_\infty = \operatorname*{ess\,sup}_{s \in S} |f(s)| = \inf\{x \in \mathbb{R} : \mu(s : |f(s)| > x) = 0\}. \tag{2.5}$$

This definition of $\mathbb{L}^p$-spaces leads to the results that $\|.\|_p$ is in fact a norm for all $p \geq 1$. We also find that the metric associated with this norm is indeed a metric. We will not prove these results and refer to Cai and Hsing (2015) for a proof. This means that $\mathbb{L}^p$-spaces are normed vector spaces, and the Riesz-Fischer theorem states that

Theorem 2.8. *The space $\mathbb{L}^p$ is complete for each $p \geq 1$.*

A proof for this theorem is also included in Cai and Hsing (2015). A consequence of this is that these spaces also fall under the category of Banach spaces, meaning we now have Banach spaces with functions for each $p \geq 1$. Note that these function spaces are not finite dimensional. This means that we have already defined metrics, norms and function spaces in the infinite dimensional setting.

2.1.3 Hilbert spaces

While we are finished with extending concepts of distance and vector spaces into an infinite dimensional setting, not all intuitive aspects of finite dimensions carry over yet. One such aspects that is yet to carry over is that of orthogonality. Having a concept of angle between elements of a Banach space is important and to generalize this, we need a good definition of an inner product.

Definition 2.9. *An inner product on a vector space $\mathbb{V}$ is a function $\langle ., . \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ that satisfies*

1. *$\langle v, v \rangle \geq 0$, and $\langle v, v \rangle = 0$ if and only if $v = 0$,*
2. *$\langle a_1 v_1 + a_2 v_2, v \rangle = a_1 \langle v_1, v \rangle + a_2 \langle v_2, v \rangle$, and*
3. *$\langle v_1, v_2 \rangle = \langle v_2, v_1 \rangle$,*

for every $a_1, a_2 \in \mathbb{R}$ and $v_1, v_2, v \in \mathbb{V}$.

Note that this definition is only for real vector spaces, since those are usually what is encountered in fda.

A vector space with an inner product $(\mathbb{V}, \langle ., . \rangle)$is called an inner product space. We can connect inner product spaces with normed spaces and metric spaces through the following theorem:

Theorem 2.10. *An inner product space $\mathbb{V}$ induces a norm, called its **canonical norm**, defined by*

$$\|v\| := \sqrt{\langle v, v \rangle}, \tag{2.6}$$

for $v \in \mathbb{V}$. With this norm, every inner product space is also a normed vector space and has all the properties of a normed space. In particular, it also follows the Cauchy-Schwarz inequality

$$|\langle v_1, v_2 \rangle| \leq \|v_1\| \|v_2\|, \tag{2.7}$$

for $v_1, v_2 \in \mathbb{V}$.

Note that while every inner product naturally induces a canonical norm and is a normed space with its canonical norm, this does not work the other way around. Not every norm is induced by an inner product, and therefore not every normed space is an inner product space.

Definition 2.11. *An inner product space that is complete under the metric associated with its canonical norm, induced by the inner product, is called a **Hilbert** space.*

Hilbert spaces are very important, since they carry all the important concepts of geometry that we naturally expect from spaces. For example, any finite dimensional inner product space is a Hilbert space. This includes all $\mathbb{R}^n$-spaces with the standard Euclidean inner product. The concept of a Hilbert space creates a framework that carries over most of these central concepts into an infinite dimensional setting. Since Hilbert spaces have inner products, they also have projections, which may not be intuitive in an infinite dimensional setting. Another important element that we will use from this, is the generalization of eigenvalue decomposition into the infinite dimensional setting.

As mentioned, not all normed vector spaces, and therefore not all Banach spaces, are Hilbert spaces. In fact, it can be hard to find Hilbert spaces in the infinite dimensional setting. Previously, we discussed that $\mathbb{L}^p(S, \mathscr{A}, \nu)$ was a Banach space for $p \geq 1$. These are not all Hilbert spaces, but

Theorem 2.12. *The Banach space $\mathbb{L}^2(S, \mathscr{A}, \nu)$, with the inner product*

$$\langle f_1, f_2 \rangle = \int_S f_1 f_2 d\nu, \tag{2.8}$$

for $f_1, f_2 \in \mathbb{L}^p(S, \mathscr{A}, \nu)$, is a Hilbert space.

This can clearly be seen, as the inner product induces the canonical norm that is equivalent to the 2-norm as defined in equation (2.4). Technically, the $\mathbb{L}^2$-space is not one of functions, but of equivalence classes of functions. Two functions are considered equivalent if they are the same almost everywhere, meaning they only differ on a set with measure zero. For simplicity, we will simply refer to this space as a Hilbert space of functions.

The $\mathbb{L}^2$-space is very important for fda, as this is the Hilbert space most functions will exist in. In particular, we will work with the $\mathbb{L}^2$-space on the measure space $([0,1], \mathscr{B}, \mu)$, where $[0,1]$ is the support of the functions, $\mathscr{B}$ is the Borel σ-algebra on $[0,1]$ and μ is the Lebesgue measure. This Hilbert space is often referred to as the $\mathbb{L}^2[0,1]$. From now onwards, when referring to a Hilbert space $\mathbb{H}$ of functions, a space such as the $\mathbb{L}^2$-space of equivalence classes is meant.

2.1.4 Hilbert-Schmidt spaces

A final space that needs introduction is a space of operators. In the next section, we will discuss the properties of operators in more details, but the general space some of them exist in will be introduced here. Operators we consider in this so-called Hilbert-Schmidt space are functions $A : \mathbb{H} \to \mathbb{H}$ that act on a Hilbert space and have the finite Hilbert-Schmidt norm

$$\|A\|_{HS}^2 := \sum_{i \in I} \|Ae_i\|^2, \tag{2.9}$$

where $\|.\|$ is the canonical norm on the Hilbert space $\mathbb{H}$ induced by the inner product and $\|.\|_{HS}$ is called the Hilbert-Schmidt (HS) norm. The set $\{e_i\}_{i \in I}$ is a complete orthonormal basis for the Hilbert space. An operator $A : \mathbb{H} \to \mathbb{H}$ is called a Hilbert-Schmidt operator when $\|A\|_{HS} < \infty$.

In practice, we will often work with finite dimensional representations of functional data. In these cases, the Hilbert-Schmidt norm is often called the Frobenius norm. Generally, the

Frobenius norm is defined for matrices instead of operators and is defined as.

$$\|A\|_F := \sqrt{\mathrm{Tr}\,(A^*A)}, \tag{2.10}$$

where A^* is the adjoint of A, as defined by theorem 2.17. When discussing theoretical properties of covariance operators, the Hilbert-Schmidt norm will be used. For practical applications when working with finite dimensional representations of functions, the Frobenius norm will be applied.

One important properties of Hilbert-Schmidt operators is that the product of two Hilbert-Schmidt operators has finite trace-class norm. The trace-class norm on an operator A is defined as

$$\|A\|_{Tr} := \mathrm{Tr}\,(|A|) := \mathrm{Tr}\left(\sqrt{A^*A}\right), \tag{2.11}$$

where,

$$\mathrm{Tr}\,(A) = \sum_{i\in I}\langle Ae_i, e_i\rangle. \tag{2.12}$$

Since the product of two Hilbert-Schmidt operators is has finite trace-class norm, the Hilbert-Schmidt inner product can be defined as:

$$\langle A, B\rangle_{HS} = \mathrm{Tr}\,(A^*B) = \sum_{i\in I}\langle Ae_i, Be_i\rangle, \tag{2.13}$$

where the second inner product is the inner product on the Hilbert space. The norm associated with the Hilbert-Schmidt inner product is the Hilbert-Schmidt norm. With this inner product and its associated norm, the space of Hilbert-Schmidt operators is also a Hilbert space. This space can be shown to be isomorphic to

$$\mathbb{H}^* \otimes \mathbb{H}, \tag{2.14}$$

the tensor product (definition 2.23) of Hilbert spaces, where $\mathbb{H}^*$ is the dual space of $\mathbb{H}$.

As mentioned, this is the space where the covariance operators exist in. Since we have mentioned this term quite a few times now, it seems prudent to define exactly what we mean by it.

Definition 2.13 (Covariance operator)**.** *A covariance operator is a linear operator which is non-negative, self-adjoint and trace-class.*

2.2 Operators in Hilbert spaces

Functions live in function spaces. In particular, the kinds of functions that are studied in fda typically live in Hilbert spaces. As with most fields of mathematics, we are not as interested in the contents of these spaces as we are in what we can do with these spaces. Specifically, the objects that we want to study are functions that act on Hilbert spaces. These are often called operators or functionals and are widely studied in the field of functional analysis. In this section, we will explore operators on normed spaces in general and Hilbert space in particular.

In addition, we will also look at linear functionals, which are a subset of operators that have certain interesting properties.

One other thing to note is the use of the word 'functional' in this context. This refers to functions between normed spaces, and it is unrelated to the use of the word in the context of 'functional data'. Note this difference to avoid confusion. This section will mostly concern 'functionals' in the context of functions that work on normed spaces, not the data.

2.2.1 Linear operators

First, we will investigate the properties of linear operators on normed spaces. First, we need some definitions:

Definition 2.14. *Given normed spaces $\mathbb{V}_1$ and $\mathbb{V}_2$, a linear operator is a function $A : \mathbb{V}_1 \to \mathbb{V}_2$ with the following property:*

$$A(\alpha v + \beta w) = \alpha Av + \beta Aw, \tag{2.15}$$

for $\alpha, \beta \in \mathbb{R}$ and $v, w \in \mathbb{V}_1$.

Take note of the notation Av in this context, which is an often used shorthand for $A(v)$, the result of applying a function A to an element v.

A restriction that is often placed on operators, both linear and in general, is that they are bounded:

Definition 2.15. *Given $\mathbb{V}_1, \mathbb{V}_2$ normed spaces with norms $\|.\|_1$, $\|.\|_2$ respectively and a linear operator $A : \mathbb{V}_1 \to \mathbb{V}_2$. A is bounded if there exists a $0 < C < \infty$, such that*

$$\|Av\|_2 \leq C\|v\|_1, \tag{2.16}$$

for all $v \in \mathbb{V}_1$.

This is quite a restrictive definition. For all elements of a vector space, there is one finite value C which is the upper bound of the increase in norm due to transformation. In fact, bounded linear operators between normed spaces are uniformly continuous. This relation goes both ways, so boundedness and uniform continuity are equivalent concepts for operators. We often refer to the space of bounded operators between normed spaces $\mathbb{V}_1$ and $\mathbb{V}_2$ as $\mathscr{B}(\mathbb{V}_1, \mathbb{V}_2)$. This vector space becomes a normed vector space under the *operator norm*

$$\|A\| = \sup_{v \in \mathbb{V}_1,\ \|v\|_1 = 1} \|Av\|_2. \tag{2.17}$$

In writing, this means that the operator norm of A is the supremum over the $\|.\|_2$ over all elements in the unit circle of $\mathbb{V}_1$ under the $\|.\|_1$. This space of bounded operators has some interesting properties, such as:

Theorem 2.16. *Given $\mathbb{V}_1$ and $\mathbb{V}_2$, normed linear spaces and $\mathbb{V}_2$ complete. Then $\mathscr{B}(\mathbb{V}_1, \mathbb{V}_2)$ with the operator norm 2.17 is a Banach space.*

Since Hilbert spaces are also Banach spaces, the space of bounded operators between Hilbert spaces is a Banach space, and therefore complete. There are several relevant properties that we can apply to the space of bounded operators between Hilbert spaces, that are generalizations of concepts in finite dimensions. The first one of these, is the generalization of the concept of the matrix transpose. For matrices, we can simply find the transpose by flipping it over the diagonal, but this is not something that can be done in infinite dimensions. We define the related concept of an *adjoint* operator based on the property

Theorem 2.17. *Let $\mathbb{H}_1$ and $\mathbb{H}_2$ be Hilbert spaces with inner products $\langle .,. \rangle_1$ and $\langle .,. \rangle_2$. For every $A \in \mathscr{B}(\mathbb{H}_1, \mathbb{H}_2)$, there exists a unique $A^* \in \mathscr{B}(\mathbb{H}_2, \mathbb{H}_1)$ such that*

$$\langle Ax_1, x_2 \rangle_2 = \langle x_1, A^* x_2 \rangle_1, \tag{2.18}$$

for all $x_1 \in \mathbb{H}_1$ and $x_2 \in \mathbb{H}_2$.

When we have $\mathbb{H}_1 = \mathbb{H}_2$, A is called self-adjoint when $A^* = A$. There are several more important properties of operators, and their adjoints, on Hilbert spaces:

Theorem 2.18. *Let $A, B \in \mathscr{B}(\mathbb{H}_1, \mathbb{H}_2)$ be a bounded operator on Hilbert spaces. Then,*

1. $(A^*)^* = A$,
2. $\|A^*\| = \|A\|$,
3. $\|A^* A\| = \|A\|^2$,
4. $(AB)^* = B^* A^*$,
5. $\mathrm{Ker}(A) = (\mathrm{Im}(A^*))^\perp$, *and*
6. $\mathrm{rank}(A^*) = \mathrm{rank}(A)$.

Further, we find that covariance operators are all trace-class. This means that a covariance operator A has a finite trace-class norm. In other words, using the definition of the trace of A from equation (2.12), we find that $\mathrm{Tr}\,(A) < \infty$. As mentioned in that same paragraph, this also implies that A is an element of the Hilbert-Schmidt space of trace-class operators. This means that all covariance operators exist within the Hilbert-Schmidt space.

Covariance operators are also positive semidefinite. We call an operator $A \in \mathscr{B}(\mathbb{H})$ positive semidefinite when for all $x \in \mathbb{H}$, $\langle Ax, x \rangle \geq 0$. When A is positive semidefinite, we write $A \geq 0$. Similarly, $A \in \mathscr{B}(\mathbb{H})$ is positive definite when $\langle Ax, x \rangle > 0$ for all $x \in \mathbb{H}$. We write $A > 0$ and call A positive.

One critical property of self-adjoint positive semidefinite operators is that they allow a type of square root.

Theorem 2.19. *Let $A \in \mathscr{B}(\mathbb{H})$ for a Hilbert space $\mathbb{H}$. If A is positive semidefinite, there is a unique positive semidefinite operator $B \in \mathscr{B}(\mathbb{H})$ that satisfies $B^2 = A$ and commutes with any operator that commutes with A.*

In practice, this element $B \in \mathscr{B}(\mathbb{H})$ is called $A^{\frac{1}{2}}$. This unique square root is also self-adjoint. The fact that the square root of a positive semidefinite self-adjoint commutes with all the elements its square commutes with, is important when considering the following theorem:

Theorem 2.20. *Let $A, B \in \mathscr{B}(\mathbb{H})$ be two self-adjoint, positive semidefinite operators on a Hilbert space. If $AB = BA$, then $(AB)^* = AB$, or AB is also self-adjoint.*

This follows from theorem 2.18, property 4, by taking the adjoints from both A and B on the right-hand side and applying the commutation.

2.2.2 Linear functionals

A linear functional on a Hilbert space $\mathbb{H}$ is a bounded linear operator $\ell : \mathbb{H} \to \mathbb{R}$. The space $\mathscr{B}(\mathbb{H}, \mathbb{R})$ these functionals live in, is also called the *dual space* of $\mathbb{H}$. This dual space has a specific form as a result of the Riesz Representation Theorem:

Theorem 2.21. *Let $\mathbb{H}$ be a Hilbert space with inner product $\langle ., . \rangle$ and norm $\|.\|$ and let $\ell \in \mathscr{B}(\mathbb{H}, \mathbb{R})$. There exists a unique $e_\ell \in \mathbb{H}$ called the representer of f, such that*

$$Lx = \langle x, e_\ell \rangle, \tag{2.19}$$

for all $x \in \mathbb{H}$ and $\|\ell\| = \|e_\ell\|$.

One of the consequences of this theorem is that the dual space of $\mathbb{H}$ has the same structure as $\mathbb{H}$ itself. The dual space is isomorphic to $\mathbb{H}$, which is easy to prove by using the above theorem to construct a bijection. In fact, the relationship is even stronger, since it is an *isometric isomorphism*. This means that distances are also preserved by the bijection between these spaces.

2.3 Functional data analysis

So far, some theory about both linear operators and linear functionals has been discussed. Both of these are relevant for studying functional data. To start, we need a clear definition of what functional data is. It has already been mentioned that the functions that are studied live in a Hilbert space. An element $x \in \mathbb{H}$ is something deterministic, while in statistics the objects of study are generally stochastic. In that vein, need a way to randomly select one function from all possible functions in our space, where the choice of function is guided by our probability distribution $\mathbb{P}$. Similar to how random variables are functions from a probability space $(\Omega, \mathscr{F}, \mathbb{P}) \to \mathbb{R}$, a random process can be defined as

$$X : (\Omega, \mathscr{F}, \mathbb{P}) \to \mathbb{H}. \tag{2.20}$$

This means that for some $\omega \in \Omega$, $X(\omega) \in \mathbb{H}$. Often, the argument is dropped and $X \in \mathbb{H}$ will refer to a random element in $\mathbb{H}$. Now that a definition of functional data is established, other metrics for these random processes can be introduced, starting with the notion of a mean.

Definition 2.22. *Given X a random element of $\mathbb{H}$ and $\mathbb{E}[\|X\|] < \infty$, the mean of X is defined as the Bochner integral*

$$\mu = \mathbb{E}[X] := \int_\Omega X d\mathbb{P}. \tag{2.21}$$

This is a natural extension of the concept of the mean to the functional data. Similar to the mean for a random variable, it is a sum of all possible values of the random element, weighted by the probability of those values. Here, it is a weighted realization of all possible elements of $\mathbb{H}$, so another element of $\mathbb{H}$ is returned. This other element is not random, just like the mean in the finite dimensional case, it is a property of the space.

An alternative definition of the mean can be found using the Riesz representation theorem (2.21). The assumption that the expected value of the norm of X is finite can also be used to define a functional $\ell : \mathbb{H} \to \mathbb{R}$ such that

$$\ell(f) = \mathbb{E}\left[\langle X, f\rangle\right] = \langle f, e_\ell\rangle, \tag{2.22}$$

where we can define μ as the representer e_ℓ.

Now that we have a definition of the mean for functional data, it can be used to define the variance and covariance operators. The focus will purely be on the covariance. In the finite dimensional case, the autocovariance matrix is defined as

$$\mathcal{K}_{XX} = \mathbb{E}\left[(X - \mathbb{E}X)(X - \mathbb{E}X)^\perp\right] = \mathbb{E}\left[(X - \mathbb{E}X) \otimes (X - \mathbb{E}X)\right], \tag{2.23}$$

where the $\otimes$ operator is known as the tensor product, which is defined as:

Definition 2.23. *Let X_1, X_2 be elements of Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$. The tensor product operator $(X_1 \otimes_1 X_2) : \mathbb{H}_1 \to \mathbb{H}_2$ is defined as*

$$(X_1 \otimes_1 X_2)Y = \langle X_1, Y\rangle_1 X_2, \tag{2.24}$$

for $Y \in \mathbb{H}_1$. When $\mathbb{H}_1 = \mathbb{H}_2$, we drop the index numbers and use $\otimes$ instead.

Similarly to (2.23), but for random elements with $\mathbb{E}\left[\|X\|\right] < \infty$, we define the covariance operator as follows:

Definition 2.24. *Let X be a random element of $\mathbb{H}$ with $\mathbb{E}\left[\|X\|\right] < \infty$. The covariance operator $\mathscr{K}$ is a function $\mathbb{H} \times \mathbb{H} \to \mathbb{H}$ given by the Bochner integral*

$$\mathscr{K} = \mathbb{E}\left[(X - \mu) \otimes (X - \mu)\right] := \int_\Omega (X - \mu) \otimes (X - \mu) d\mathbb{P}, \tag{2.25}$$

and $\mathscr{K} \in \mathscr{B}_{HS}(\mathbb{H})$ is a Hilbert-Schmidt operator (2.9).

Another way to represent this result, that is generally more familiar, is

$$\mathbb{E}\left[(X - \mu) \otimes (X - \mu)\right] = \mathbb{E}\left[X \otimes X\right] - \mu \otimes \mu. \tag{2.26}$$

This representation is much easier to work with, especially when $\mu = 0$. From this point, it will be assumed that $\mu = 0$. If $\mu \neq 0$, then the random process can simply be replaced with $\hat{X} = X - \mu$, which is also in $\mathbb{H}$ and does have mean zero. With $\mu = 0$, the integral simplifies to

$$\mathscr{K} = \mathbb{E}\left[X \otimes X\right] =: \int_\Omega X \otimes X d\mathbb{P}. \tag{2.27}$$

With the assumption that $\mu = 0$, our covariance operator is a positive semi-definite, trace-class operator. Trace-class means that the operator is finite under the trace-class norm

$$\|\mathscr{K}\|_{Tr} = \mathrm{Tr}\left(\sqrt{\mathscr{K}^*\mathscr{K}}\right) = \mathrm{Tr}\left(|\mathscr{K}|\right), \tag{2.28}$$

since $\mathscr{K}^* = \mathscr{K}$ because covariance operators are self-adjoint. Another representation of this norm is

$$\|\mathscr{K}\|_{Tr} = \sum_{i=1}^{\infty}\langle \mathscr{K} e_i, e_i\rangle = \sum_{i=1}^{\infty}\langle X, e_i\rangle^2 = \mathbb{E}\|X\|^2, \tag{2.29}$$

where $\{e_i\}$ is any complete orthonormal basis for $\mathbb{H}$. The above statement holds because $\langle \mathscr{K} f, g\rangle = \mathbb{E}\left[\langle x, f\rangle\langle X, g\rangle\right]$ for any $f, g \in \mathbb{H}$. Since this sum is over an infinite basis, the trace-class property means that the individual terms of the sum go to zero 'very fast'. Because of this fast-decaying property, the operator admits an eigenvalue decomposition

$$\mathscr{K} = \sum_{i=1}^{\infty}\lambda_i e_i \otimes e_i, \tag{2.30}$$

where the $\{e_i\}$ are a complete orthonormal basis. Because of the trace-class property, the sequence $\{\lambda_i\}$ either has finite non-zero elements or it converges to zero 'fast enough'.

So far this has mostly been approached from an autocovariance perspective, but it is also usefull to specify the definition for cross-covariance between two random processes that possibly have their realizations in different Hilbert spaces. Given two random elements X_1, X_2 from $\mathbb{H}_1, \mathbb{H}_2$ respectively, both with mean zero. Then the cross-covariance is

$$\mathscr{K}_{1,2} = \int_{\Omega} X_2 \otimes_2 X_1 d\mathbb{P}, \tag{2.31}$$

if $\|X_1\|_1^2, \|X_2\|_2^2 < \infty$. This operator is once again an element of a Hilbert-Schmidt space $\mathscr{B}_{HS}(\mathbb{H}_2, \mathbb{H}_1)$. The adjoint of $\mathscr{K}_{1,2}$ is $\mathscr{K}_{2,1} = \int_{\Omega} X_1 \otimes_1 X_2 d\mathbb{P}$.

2.4 Estimators for fda

There are estimators that are currently used in fda. Most of these are based on estimators that were originally defined on finite dimensional $\mathbb{R}^p$ Hilbert spaces and then adapted and extended to the infinite dimensional setting where possible. Many of these estimators are based on reducing the infinite dimensionality of functional data to some finite dimensional space. These methods are often based on the *Karhunen-Loève* theorem (Karhunen, 1947; Loève, 1946) that represents a random element X of a Hilbert space of functions $\mathbb{H}$ as an infinite linear combination of orthonormal basis functions.

Theorem 2.25 (Karhunen-Loève)**.** *Let $\{X(t) : t \in T\}$ be a mean-square stochastic process with mean zero, defined over a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then,*

$$\lim_{n\to\infty}\sup_{t\in T}\mathbb{E}\left[X(t) - X_n(t)\right]^2 = 0, \tag{2.32}$$

where

$$X_n(t) := \sum_{i=1}^{n} I_i e_i(t), \tag{2.33}$$

where

$$I_i = \int_0^1 X(t)e_i(t)dt. \tag{2.34}$$

What this theorem states is that the representation (2.33) converges to the actual function in mean-square. This means that by increasing n, one can get arbitrarily close to the function X. With a certain accepted error, this can create a finite dimensional representation that approaches the function to a chosen degree of error. This theorem can be used as an infinite dimensional version of principal component analysis. It is mostly for the mathematical foundations of the theory, but cannot be directly seen in the practical implementation of the distance functions.

2.4.1 Frobenius distance

The Frobenius norm was mentioned earlier as a finite dimensional version of the Hilbert-Schmidt norm (2.10). Since our representations of functional data are finite dimensional, we can use this norm to define the Frobenius metric

Definition 2.26. *For two matrices A and B, the Frobenius metric is defined as*

$$d_F(A,B) := \|A-B\|_F = \sqrt{Tr\left[(A-B)^*(A-B)\right]}. \tag{2.35}$$

This is a very simple metric that can be used to define the distance between two matrices. It scales perfectly to the infinite dimensional case of covariance operators, under its alternative name of the Hilbert-Schmidt norm and associated metric. This makes it a great distance function as a test statistic. It is also known as the Euclidean distance between matrices.

One downside to the use of this metric, especially when using it as the distance function for a least-squares estimator, is that it does not respect the geometry of the space of covariances. Which do not exist on a Euclidean space. (Dryden et al., 2009; Pigoli et al., 2014)

2.4.2 log-Euclidean distance

The first example of a non-Euclidean distance function can be found through the use of matrix logarithms. Taking the logarithm of a matrix may not seem simple, so let us define this action:

Definition 2.27. *Given a matrix S, the logarithm of S is given as:*

$$\log S = U \log \Lambda U^*, \tag{2.36}$$

where U is an orthonormal matrix with Eigenvectors and Λ is a diagonal matrix with Eigenvalues on its diagonal. $S = U\Lambda U^*$ is known as the spectral or Eigenvalue decomposition of S. The matrix $\log\Lambda$ consists of the logarithms of the Eigenvalues on the diagonal.

Using the matrix logarithm, we can adapt the Frobenius, or Euclidean, distance with the logarithm turn it non-Euclidean. This leads to the definition of the log-Euclidean distance, as introduced by Arsigny et al. (2007).

Definition 2.28. *Given two positive definite matrices A and B, we define the log-Euclidean distance between A and B as*

$$d(A,B)_{LE} := \|\log(A)-\log(B)\|_F = \sqrt{Tr\left[(\log(A)-\log(B))^*(\log(A)-\log(B))\right]}. \tag{2.37}$$

This metric can also be used to generate a test statistic for the distance between matrices. It respects the geometry of the space more, since it's no longer a Euclidean metric. However, this does come at a cost of computation time.

2.4.3 Procrustes size-and-shape metric

The Procrustes size-and-shape metric is a measure of dissimilarity between two covariance matrices. It aims to measure the dissimilarity between two covariance matrices while accounting for both rotation and scaling differences. The definition suggested here does not include translation, which is irrelevant since we consider our data to have mean zero. As a result, if two covariance matrices have the same structure and variability but are simply rotated or scaled differently, the Procrustes alignment procedure would find an optimal transformation that aligns them perfectly, resulting in a Procrustes distance of zero.

We look to Pigoli et al. (2014) and Masarotto et al. (2019) for the definition of the Procrustes metric:

Definition 2.29. *Given two positive definite matrices A and B, the Procrustes size-and-shape distance between them is defined as:*

$$d(A,B)_{Pr} := \inf_{U:U^*U=I} \|A^{\frac{1}{2}} - B^{\frac{1}{2}}U\|_F, \tag{2.38}$$

where U is a unitary operator.

2.4.4 Riemannian distance

We also apply the Riemannian distance function. This metric is based on the properties of the geometry of Riemannian manifolds and can also be explained as being the geodesic path length distance function. This distance function will be explained in more detail further ahead, under the name of the geodesic path length distance metric in definition 3.1 and section 3.4.

2.5 Stochastic processes

Because of the spectral theorem, a result of the Karhunen-Loéve theorem (2.25), the covariance operator can be written as

$$\mathscr{K} = \sum_{i=1} \lambda_i e_i \otimes e_i, \tag{2.39}$$

where $\{\lambda_i\}$ is the set of eigenvalues of $\mathscr{K}$. This eigenvalue decomposition then leads to the following expression for a random process χ:

$$\chi = \mu + \sum_{i=1} Z_i e_i, \tag{2.40}$$

for $\mu = \mathbb{E}[\chi]$ and $Z_i = \langle \chi - \mu, e_i \rangle$ are uncorrelated random variables with mean zero. Further, we have that $\mathrm{Var}(Z_i) = \lambda_i$. The Z_i are often referred to as the principal components of χ.

For our simulations, we want to simulate directly from the covariance operator, using random uncorrelated random variables for each eigenvector. Define $\varepsilon_i := Ze_i$, where $\{e_i\}$ is an

orthonormal set and $Z \sim N(0,1)$. This gives us the set $\{\varepsilon_i\}$, which are uncorrelated random variables with mean 0. Then we can define our random process as

$$\chi = \mu + \sum_{i=1} \mathscr{K} \varepsilon_i. \tag{2.41}$$

Chapter 3

Trace-class extended Hilbert-Schmidt space

Many of the estimators that exist for covariances, including the ones we have discussed in section 2.4, are based on the mathematics of the Karhunen-Loève theorem. Using this theorem turns the infinite dimensional nature of the data into some finite dimensional representation. One example of this application is Lawson and Lim (2013b), where they use the finite dimensionality to add scalar elements to the covariances. Using the canonical metric on the Hilbert-Schmidt space does give an estimator that respects most of the infinite dimensional properties of covariance operators. This metric does, generally, not respect the geometry of the set of positive operators within the space. Using the canonical metric to define a mean element, does not necessarily return a covariance, which may not be desirable depending on the application. This is because the covariance operators don't exist in a linear subspace of the Hilbert-Schmidt space. They exist on a set that can be understood as an extension of the cone that positive matrices live on in Hilbert spaces of matrices. Similarly, covariance operators live in a non-linear subset of the Hilbert-Schmidt space.

One way to understand that the mean of two covariances may not be a covariance, is by thinking about the geometry of earth. The middle point of two points on the surface of the earth is always somewhere below the earth's surface when considering the distance induced by the Euclidean distance in three dimensions. The shortest line between two points goes through the inner areas of the earth. For applications such as determining the distance between two places, a more useful metric of distance in practice is the distance over the surface of the earth. To find these shortest paths, the concept of geodesics, the shortest curve across a surface, is deployed.

In this chapter, we will introduce the space that covariance operators exist in and mention the vital properties once again. Then, on the set of covariance operators, we will define an inner product and use its canonical norm to induce a metric on the set of covariance operators.

3.1 Extended Hilbert-Schmidt algebra

Let $\mathscr{B}(\mathbb{H})$ be the Hilbert space of bounded operators working on a Hilbert space $\mathbb{H}$. Within this space, we find the subspace of Hilbert-Schmidt operators mentioned earlier. Algebraically, this space can also be represented as the bilateral ideal of Hilbert-Schmidt operators $HS(\mathbb{H})$, the algebra of bounded linear operators on a general complex Hilbert space $\mathbb{H}$ (Larotonda, 2008; Lawson & Lim, 2013a). Since $\mathscr{B}_{HS}(\mathbb{H})$, the space of Hilbert-Schmidt operators, is a Banach space, it also holds that $HS(\mathbb{H})$ is a Banach algebra. Within our space of bounded operators ($\mathscr{B}(\mathbb{H})$), we define the extended Hilbert-Schmidt algebra

$$\mathscr{H}_{\mathbb{C}} := \{A + \lambda I : A \in HS(\mathbb{H}),\ \lambda \in \mathbb{C}\}, \tag{3.1}$$

which is a complex linear subalgebra. When we define the inner product on this space, such that Hilbert-Schmidt operators are orthogonal to scalar operators (λI), this space becomes a Hilbert space. This is achieved through the inner product

$$\langle A + \lambda I, B + \nu I\rangle_{\mathscr{H}} := \mathrm{Tr}\,(A^* B) + \lambda\bar{\nu}. \tag{3.2}$$

However, since we are dealing with functional data that consists of real functions, we will focus on the real part of $\mathscr{H}_{\mathbb{C}}$. For operators, being real is the equivalent to self-adjoint, which is another condition for our space of covariance operators. This real part of $\mathscr{H}_{\mathbb{C}}$ is defined as

$$\mathscr{H}_{\mathbb{R}} := \{A + \lambda I : A^* = A,\ A \in HS(\mathbb{H}),\ \lambda \in \mathbb{R}\}, \tag{3.3}$$

which is also a Hilbert space with the inner product defined in (3.2) without the need for the conjugate of ν. The subset of positive elements of $\mathscr{H}_{\mathbb{R}}$ is equal to the space of covariance operators, extended with positive $\lambda \in \mathbb{R}_{>0}$. Call this subset of positive elements of the real extended Hilbert-Schmidt space $\mathscr{P}$. Based on this extension of Hilbert-Schmidt operators, Larotonda, 2008 shows that this space is a Riemannian manifold.

Definition 3.1. *A Riemannian manifold $(\mathcal{M}, g)$ is a real smooth manifold $\mathcal{M}$ with positive definite inner product g_p defined on the tangent space $T_p\mathcal{M}$ at each $p \in \mathcal{M}$. The inner products g_p are called the Riemannian metric.*

Where the Riemannian metric for $\mathscr{P}$, for all $p \in \mathscr{P}$ is defined as:

$$\langle x, y\rangle_p := \langle p^{-1}x, yp^{-1}\rangle_{\mathscr{H}} = \langle xp^{-1}, p^{-1}y\rangle_{\mathscr{H}}, \qquad \text{and} \tag{3.4}$$

$$\|x\|_p := \langle x, x\rangle_p^{\frac{1}{2}} = \|p^{-\frac{1}{2}} x p^{-\frac{1}{2}}\|_{\mathscr{H}}. \tag{3.5}$$

This then leads to the definition of a metric based on geodesics, the shortest path between two points over the manifold. For $p, q \in \mathscr{P}$, these geodesics are defined as $\gamma_{pq}(t) := p^{\frac{1}{2}}(p^{-\frac{1}{2}} q p^{-\frac{1}{2}})^t p^{\frac{1}{2}}$ for $t \in [0, 1]$. Note that, $\gamma_{pq}(0) = p$ and $\gamma_{pq}(1) = q$. The distance between $p, q \in \mathscr{P}$ can then be defined as the path length of γ. Path length is defined as

$$L(\gamma) := \int_0^1 \|\gamma'(t)\|_{\gamma(t)} dt, \tag{3.6}$$

where $\|.\|_{\gamma(t)}$ is the Riemannian norm at the point $\gamma(t)$ and $\gamma'(t)$ is the tangent vector of γ at t and is itself an element of $T_{\gamma(t)}\mathcal{M}$. This path length can be used to define a metric on the tangent space. As found in Larotonda, 2008, this metric is

$$d(p,q) := L(\gamma_{pq}) = \|\ln p^{\frac{1}{2}} q^{-1} p^{\frac{1}{2}}\|_{\mathscr{H}}. \tag{3.7}$$

This metric was adopted as the Thompson metric and applied to covariance matrices in Lawson and Lim, 2013b. In particular, they showed that this the mean attained with the use of this metric is a global minimum for sum of squared distances.

The issue with this application of means to the extended Hilbert-Schmidt space, is that the positive elements of the real part of this extension are not Hilbert-Schmidt operators themselves. By adding an identity element, which does not have a finite trace, ensures that none of the elements of $\mathscr{P}$ are trace-class.

3.2 Trace-class extended Hilbert-Schmidt algebra

To solve the issue that these elements themselves are not trace-class operators, we will propose a different extension. Again, starting from the space of bounded operators on a Hilbert space ($\mathscr{B}(\mathbb{H})$), we identify this space with the bilateral ideal of Hilbert-Schmidt operators $HS(\mathbb{H})$. Here, $\mathbb{H}$ can be any Hilbert space, including complex ones. We define the trace-class extended Hilbert-Schmidt algebra:

$$\mathscr{H}_{\mathbb{C}} := \{A + \lambda\Sigma : A \in HS(\mathbb{H}),\ \lambda \in \mathbb{C}\}, \tag{3.8}$$

where Σ could be any non-zero trace-class operator. For the simulation, we use

$$\Sigma := \sum_{n=1} n^{-4}[\cos 2\pi nt \otimes \cos 2\pi nt]. \tag{3.9}$$

For the inner product on this space, the definition should be similar to that of the regular extended Hilbert-Schmidt algebra, where trace-class operators are orthogonal to scalar operators. In this case, the inner product will be defined such that a diagonal trace-class operator is orthogonal to a positive semidefinite singular operator. In order to use that definition of orthogonality, the following singularity decomposition is applied:

Definition 3.2. *For $A \in HS(\mathbb{H})$, we define the singularity decomposition for a given non-zero, diagonal trace-class operator Σ as*

$$\bar{A} + \alpha_A \Sigma, \tag{3.10}$$

where $\alpha_A \in \mathbb{R}$ is defined such that the spectral decomposition for $\bar{A} = A - \alpha_A\Sigma := \sum_{n=1}^{\infty} \nu_n e_n \otimes e_n$ for a complete orthonormal system $\{e_n\}$ with $\{\nu_n\}$ has $\nu_i = 0$ for any i. We call $\bar{A}$ the singular element of A and α_A the singularity scalar.

What this singularity decomposition achieves is that it turns every $A \in HS(\mathbb{H})$, which may or may not be singular, into a singular operator and a trace class element. Note that if A is already singular, then $\lambda_A = 0$. With the singularity decomposition defined, this can be used in

the definition of the inner product. With the following inner product, $\mathscr{H}_{\mathbb{C}}$ becomes a Hilbert space:

$$\langle A + \lambda\Sigma, B + \nu\Sigma\rangle_{\mathscr{P}} := \mathrm{Tr}\left(\bar{A}^*\bar{B}\right) + (\alpha_A + \lambda)(\alpha_B + \nu)\mathrm{Tr}\left(\Sigma^*\Sigma\right), \tag{3.11}$$

where $\mathrm{Tr}\left(\Sigma^*\Sigma\right) = \sum_{n=1} n^{-4} = \frac{\pi^8}{9450}$ for the choice of Σ from (3.9). Again, we only care about the self-adjoint real part of $\mathscr{H}_{\mathbb{C}}$

$$\mathscr{H}_{\mathbb{R}} := \{A + \lambda\Sigma : A^* = A,\ A \in HS(\mathbb{H}),\ \lambda \in \mathbb{R}_{>0}\}. \tag{3.12}$$

Finally we define $\mathscr{P}$ as the subset of $\mathscr{H}_{\mathbb{R}}$ of positive semidefinite operators from $HS(\mathbb{H})$, which is the same as the subset of covariance operators. For simplicity, we will simply refer to $\mathscr{P}$ as the set of extended covariance operators. The following holds for $\mathscr{P}$:

Theorem 3.3. *We define the set $\mathscr{P}$ as*

$$\mathscr{P} := \{A + \lambda\Sigma : A \in HS(\mathbb{H}),\ A^* = A,\ A \geq 0,\ \lambda \in \mathbb{R}_{>0}\}, \tag{3.13}$$

where $\Sigma = \sum_{n=1} n^{-4}[\cos 2\pi nt \otimes \cos 2\pi nt]$. For any $X, Y, Z \in \mathscr{P}$ and $f \in \mathbb{H}$, the following holds:

1. *X is positive definite*
2. *X is invertible*
3. *$X^{\frac{1}{2}} \in \mathscr{P}$*
4. *$\langle ZX, YZ\rangle_{\mathscr{P}} = \langle XZ, ZY\rangle_{\mathscr{P}}$*

Proof. We prove 1, then 2 and 3 follow trivially. X is positive definite if $\langle Xf, f\rangle > 0$. We have $\langle (A + \lambda\Sigma)f, f\rangle = \langle Af, f\rangle + \lambda\langle \Sigma f, f\rangle \geq \lambda\langle \Sigma f, f\rangle > 0$, since Σ is positive and $\lambda > 0$. From this, 2 and 3 follow, since positive operators are invertible and have a unique square root element. □

Next, for each $P \in \mathscr{P}$ we will define the inner product g_P on the tangent space $T_P\mathscr{P}$:

$$g_P(X, Y) := \langle X, Y\rangle_P := \langle P^{-1}X, YP^{-1}\rangle_{\mathscr{P}} = \langle XP^{-1}, P^{-1}Y\rangle_{\mathscr{P}}. \tag{3.14}$$

We define $\|X\|_P = \langle X, X\rangle_P^{\frac{1}{2}}$ as the norm associated with the Riemannian metric inner product. The set $\mathscr{P}$ endowed with the inner product set g_P is a Riemannian manifold.

3.2.1 Unicity

Within the extended Hilbert-Schmidt algebra, all elements are unique. This can be trivially shown, since each trace-class element is unique and by adding a scalar operator, the elements are no longer trace class and can therefore never be the same as each other. Another way to frame this, is that the function that takes an operator and scalar is injective, each pair is mapped to a unique element in the extended Hilbert-Schmidt algebra. This argument does not hold for the trace-class extended Hilbert-Schmidt space.

However, unicity of elements would be a very useful property of the trace-class extended Hilbert-Schmidt space. If there would not be unicity, then two different covariance operators may have a distance of exactly 0 within the trace-class extended Hilbert-Schmidt space and this

could cause problems with the implementation of statistics based on such a distance function. In this section, we will explore the possible unicity, or lack thereof of, of elements in the set of extended covariance operators $\mathscr{P}$.

On the basic level, elements of $\mathscr{P}$ are not unique. If $A + \lambda\Sigma \in \mathscr{P}$ for a certain $A \in HS(\mathbb{H})$, then we also have $(A + \frac{\lambda}{2}\Sigma) + \frac{\lambda}{2}\Sigma \in \mathscr{P}$ with $A + \frac{\lambda}{2}\Sigma \in HS(\mathbb{H})$, since it is still a trace-class covariance operator and $\frac{\lambda}{2} \in \mathbb{R}_{>0}$. This means that any element of $\mathscr{P}$ can be the extension of multiple different elements of $HS(\mathbb{H})$.

However, this argument against unicity requires that the two covariance operators are different on the diagonal elements only, since Σ is specified to be a diagonal operator. An equivalence relation is defined on $HS(\mathbb{H})$, where $A, B \in HS(\mathbb{H})$ are equivalent if their singular elements as defined in definition 3.2 are equal ($\bar{A} = \bar{B}$). Elements of $\mathscr{P}$ are unique with respect to the equivalence class of singular elements.

Note that the diagonal elements of a covariance operator are related to the variance of functional data. With the goal being the study of covariances of functional data, this is a compromise that can be allowed to exist in practice. We need to see through the data analysis and simulation if the theoretical problem also becomes a practical one.

3.3 Canonical metric

As mentioned earlier, Hilbert spaces are inner product spaces that are also complete metric spaces with respect to the canonical metric induced by the inner product. This canonical metric will be the first one defined and used.

For covariance operators $A, B \in HS(\mathbb{H})$, we have their extensions in $\mathscr{P}$. Based on the inner product (3.11), we define the canonical norm

$$\|A + \lambda\Sigma\|^2_{\mathscr{P}} := \langle A + \lambda\Sigma, A + \lambda\Sigma\rangle_{\mathscr{P}}. \tag{3.15}$$

This norm leads to the definition of the canonical metric on $\mathscr{P}$:

$$d_{\mathscr{P}}(A + \lambda\Sigma, B + \nu\Sigma) := \|(A + \lambda\Sigma) - (B + \nu\Sigma)\|_{\mathscr{P}}. \tag{3.16}$$

By combining both of the above definitions with (3.11), we obtain the following formula

$$\begin{aligned} d^2_{\mathscr{P}}(A + \lambda\Sigma, B + \nu\Sigma) &= \langle (A + \lambda\Sigma) - (B + \nu\Sigma), (A + \lambda\Sigma) - (B + \nu\Sigma)\rangle_{\mathscr{P}} && (3.17)\\ &= \mathrm{Tr}\left(\bar{A}^2\right) + \mathrm{Tr}\left(\bar{B}^2\right) - 2\mathrm{Tr}\left(\bar{A}\bar{B}\right) + (\lambda + \alpha_A - \nu - \alpha_B)^2\mathrm{Tr}\left(\Sigma^2\right), && (3.18)\\ &= \|\bar{A}\|^2_{HS} + \|\bar{B}\|^2_{HS} - 2\langle \bar{A}, \bar{B}\rangle_{HS} + (\lambda + \alpha_A - \nu - \alpha_B)^2\|\Sigma\|^2_{HS}. && (3.19) \end{aligned}$$

3.4 Geodesic metric

As mentioned earlier, our space $\mathscr{P}$ is a Riemannian manifold when endowed with the inner product set g_P on the tangent space $T_P\mathscr{P}$ at $P \in \mathscr{P}$ as defined in (3.14). Riemannian manifolds lend themselves to the definition of smooth functions $\gamma : [0, 1] \to \mathscr{P}$ between two elements of

within the manifold. As a smooth function, γ is differentiable and for each $t \in [0,1]$, $\gamma'(t)$ is an element of $T_\gamma(t)\mathscr{P}$, the tangent space at $\gamma(t)$. Since $\gamma'(t)$ is in the tangent space, we can measure the size of this derivative by $\|\gamma'(t)\|_{\gamma(t)}$, where $\|.\|_{\gamma(t)}$ is the norm associated with the Riemannian metric (3.4) at $\gamma(t)$.

So far, γ has only been mentioned in general as a smooth function. More specifically, γ is the shortest path between two points over the surface of the Riemannian manifold. For the extended Hilbert-Schmidt space, this is

$$\gamma_{A,B}(t) = A^{\frac{1}{2}} \left(A^{-\frac{1}{2}} B A^{-\frac{1}{2}}\right)^t A^{\frac{1}{2}}, \tag{3.20}$$

as defined by Larotonda (2008). For our trace-class extended space, we use the same definition of the geodesics.

The length of the path is based on the size of the derivative of γ, integrated over the path. This gives us the following definition for the geodesic arc length

$$L(\gamma) = \int_0^1 \|\gamma'(t)\|_{\gamma(t)} dt. \tag{3.21}$$

Based on this definition of arc length, the geodesic metric is defined as

$$d_g^2(A,B) := L(\gamma_{A,B}) = \|\ln A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}}\|_{HS}^2 = \mathrm{Tr}\left(\left(\ln A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}}\right)^* \ln A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}}\right) \tag{3.22}$$

3.5 Implementation of metrics

With the definition of the canonical and geodesic metrics on covariance operators, we have a method to determine the distance between them. For a given sample, the covariance operator is related to the variance of the basis functions, as illustrated in (2.40). Covariance operators are therefore related to their data in a similar fashion as residuals. We do not generally care about their exact values, but we do care about their distribution. Covariance operators often follow a Gaussian distribution, and that is what will be assumed here as well.

One statistic that we do care about, with regard to the covariance operator, is a concept of a mean for the distribution. For more complex distribution, we do not generally refer to these elements as means, but as centers of mass of the distribution. For a Gaussian distribution, the mean, median, mode and center of mass all coincide, but this is not true in general. The concept of the center of mass as a mean is generalized to Riemannian manifolds by the concept of the Karcher and/or Frechét mean. These names are often used intertwined, but we will define their specific uses.

Definition 3.4. *Given a complete metric space (M,d), the Karcher means for $A_{ii\in I} \subset M$ are those points $P \in M$ that locally minimize*

$$P := \underset{X\in M}{\arg\min} \sum_{i\in I} d^2(X, A_i). \tag{3.23}$$

If there is a $P \in M$ that globally minimizes the sum of squared distances, then it is called the Frechét mean.

Now that we have a functional definition of means for Riemannian manifolds, we can use this in conjunction with the two defined metrics. For both the canonical, and the geodesic metrics, we can calculate try to find the corresponding Frechét means. To do this, we need some algorithm to solve for the minimizer. Unfortunately, there is no closed-form solution to this minimizing problem. We will find the minimizers using a gradient descent algorithm. As the name implies, this requires us to find the gradients for both distance functions.

First, we need to lay some groundwork for the gradient of operators, including the derivative of the trace. As a reference, we will use the work done by Petersen and Pedersen (2008) as a basis.

Theorem 3.5. *Given a complete metric space (M, d) with $X \in M$ and F a differentiable function on X. The derivative of a trace is then defined as*

$$\frac{\partial Tr(F(X))}{\partial X} = f(X)^*, \tag{3.24}$$

and the chain rule applies as expected.

Using this definition, we can see that for $X \in \mathscr{P}$, such that $F(X) = AX$ with $A \in \mathscr{P}$, the derivative of $\mathrm{Tr}\left(F(X)\right) = \mathrm{Tr}\left(AX\right)$ is A^*.

For the canonical metric, we therefore need to determine the gradient of the sum of squared distances $\sum_i d^2(X, A_i)$, which is

$$\frac{\partial}{\partial X} \sum_{i=1} d^2_{\mathscr{P}}(X, A_i + \lambda_i \Sigma), \tag{3.25}$$

$$= \sum_{i=1} \frac{\partial}{\partial X} \left[\mathrm{Tr}\left(\bar{X}^2\right) + \mathrm{Tr}\left(\bar{A}_i\right) - 2\mathrm{Tr}\left(\bar{A}_i \bar{X}\right) + (\alpha_X - \lambda_i - \alpha_{A_i})\mathrm{Tr}\left(\Sigma^2\right) \right], \tag{3.26}$$

$$= \sum_{i=1} 2\bar{X} - 2\bar{A}_i, \tag{3.27}$$

$$= \sum_{i=1} 2(\bar{X} - \bar{A}_i). \tag{3.28}$$

Using this gradient, we can use a gradient descent algorithm to find a Karcher mean X. Since we assume that the covariances are from a Gaussian distribution and, because they are well-behaved, there should only be one Karcher mean, which is then also the Frechét mean. An important note here is that the gradient does not depend on the trace-class extension operator or the scalars in any way. This means that there is no difference between the gradient of sum of squares for the canonical operator for the trace-class extended Hilbert-Schmidt space and the scalar extended Hilbert-Schmidt space. This means that the Frechét means for both coincide.

For the geodesic metric, we can use the result from Lawson and Lim (2013a), who show that the gradient of the geodesic metric is

$$\frac{\partial}{\partial X} \sum_{i=1} d^2_g(X, A_i + \lambda_i \Sigma), \tag{3.29}$$

$$= \sum_{i=1} \frac{\partial}{\partial X} \mathrm{Tr}\left(\ln\left[X^{\frac{1}{2}} A_i^{-1} X^{\frac{1}{2}} \right] \right), \tag{3.30}$$

$$= \sum_{i=1} X^{-\frac{1}{2}} \ln\left[X^{\frac{1}{2}} A_i^{-1} X^{\frac{1}{2}} \right] X^{-\frac{1}{2}}. \tag{3.31}$$

3.5.1 Implementation issues

There were several problems that arose when attempting to implement the trace-class extension, many due to computational constraints. Firstly, many of the covariances very close to not being non-negative and rounding errors in calculations meant that in practice, the covariance operators were not considered positive definite. This problem is partly resolved with the addition of a scalar operator in the extended Hilbert-Schmidt version, while the addition of the trace-class operator does not change the matrix enough to get rid of that issue.

The second issue that appeared was caused by our definition of the inner product using the singularity decomposition from definition 3.2. By definition, this decomposition creates a matrix that is only *barely* non-negative. This again leads to an issue with positive definiteness. Solving the problem with the positive definiteness when using certain metrics requires the addition of a scalar matrix, rendering it useless to apply the trace-class extension in the first place.

Chapter 4

Data Analysis

So far, we have discussed several metrics that can be used as a distance function in a least-squares or Frechét mean and can also be used as the distance function for a test statistic. These include the Frobenius distance (section 2.4.1), Procrustes size-and-shape metric (section 2.4.3) and log-Euclidean distance (section 2.4.2). Further, we discussed the canonical metric on the extended Hilbert-Schmidt space, which is an extension of the Frobenius metric as a distance function to be used in determining a Frechét mean (section 3.3). Finally, we looked at the path-length based Geodesic distance function (section 3.4) and both the Thompson metric that it leads to on the identity-extended Hilbert-Schmidt space and the metric on the trace-class extended Hilbert-Schmidt space.

We want to investigate how these methods perform on actual data, where there is reason to believe that covariances could be a way to differentiate between samples. For this section, we use an open source dataset called `phoneme`, which is included in the R package `fda` (Ramsay et al., 2020). We will compare two samples based on their covariances and use a nonparametric statistical test to determine if they are different.

We will perform these tests with different methods of taking the mean covariance of a sample, different distance functions for the test statistic and do this for several different two-sample comparisons. The sample comparisons we look at are between the phoneme `aa` and `ao`, `aa` and `iy` and finally `aa` and `sh`. In the written order, these phonemes are increasingly different and therefore easier to tell apart.

The outcome of interest is the power of each of the two-sample permutation tests with different methods of determining the covariance of the sample and different test statistics. This could give us an idea about which method of evaluating covariances works best in a practical setting.

4.1 Data description

The `phoneme` data from the R package contains digitized speech recordings from five phonemes (`aa`, `ao`, `dcl`, `iy` and `sh`). Three of these are vowels (`aa`, `ao` and `iy`) and two consonants (`dcl` and `sh`). These phonemes are pronounced as in the following words: `aa` as in ‘dark‘; `ao` as in ‘water‘; `dcl` as in ‘dark‘; `iy` as in ‘she‘; `sh` as in ‘she‘. We expect the three vowels to be closer

in both mean structure and covariance to each other than to the consonants and among the vowels, we expect `aa` and `ao` to be closer to each other than to `iy`.

Each of the data entries is a log-periodogram. A periodogram is an estimate of the spectral density of signal at different frequencies. For the phoneme data specifically, the x-axis contains the 150 equidistant frequencies and the y-axis contains the logarithm of the estimated spectral density.

4.2 Method

As mentioned, we want to assess the power of different two-sample permutation tests. The differences are either in the choice for the 2nd sample, choice for method of determining the sample covariance mean or in the distance function that calculates the test statistic.

First up, is setting up the data for the power test. The two samples are fully loaded from the `phoneme` dataset and contain 400 functions with the log-periodogram at 150 frequencies. For calculating the covariance, it is much easier to do so with a function that has a mean of 0, so we subtract the sample mean from each of the functions first, this gives us the form as in equation (2.27). The data is very noisy and has some quite extreme fluctuations between frequencies that are very similar. This could have a big impact on the covariance of the sample. To avoid this noise dominating the covariance, we smooth the data using Fourier basis splines with 21 knots.

After smoothing the data, the data is subset. A resolution of 150 leads to covariance matrices that are 150 by 150 and this takes up unreasonable amounts of computational resources. To make the computation more rapid, we subset the data to only 50 points out of the 150 by selecting every third point from 1 to 148.

After preparing the data, we run the power test. The power test consists of 250 repeats of a permutation test with a specific method of determining the sample covariance means. For each of the repeats, the data is randomly sampled, testing either the null hypothesis or the alternative. Under the null hypothesis, both samples are taken from one of the two data populations, the choice of which population is random. Under the alternative hypothesis, both populations are randomly sampled. The sample size that we have chosen is 25 functions in each of the two samples.

We want to assess a sequence of two-sample nonparametric tests that can be used to differentiate between two samples based on their covariances. The desired outcomes that we will compare are the power of the test and the false positive rate α, based on $p < 0.05$ for a statistically significant result from the permutation test. The power is calculated as the sum of the permutation tests that returned $p < 0.05$, divided by the total number of repeats, under the alternative hypothesis. The false positive rate is calculated in the same way, but under the null hypothesis instead.

For these power tests, we will use R version 4.2.1. (R Core Team, 2021). The scripts used can all be found in on the github page (Hackmann, 2023). In order to be able to perform the required computations in this project, we have turned to Academic Leiden Interdisciplinary

Cluster Environment (ALICE) for the required computational power.

4.2.1 Covariance estimation

We will compare four different method of determining the covariance for each sample, which we refer to as the estimated sample covariance. It is interesting to see if there is an impact of the type of covariance estimation and what the size of this effect is.

The first of the estimation methods that we will use is one that will simply be referred to as the 'Sample' covariance estimation. It is named this way since it is estimated similarly to the sample variance estimation. To estimate the sample covariance, we sum the covariance matrices and then divide by the sample size.

Second, we determine what will be referred to as the 'Extended' covariance estimation. What this is, is the Frechét mean of the covariances of all functions in the sample, using the canonical metric on the trace-class extended Hilbert-Schmidt space. We find this estimator by adding diagonal trace-class matrices to each of the covariances, using random scalars that are drawn from an exponential distribution with a rate parameter of 10.

Third, we estimate with the Thompson mean. This is the Frechét mean using the geodesic path length metric on the identity-extended space. We will apply this by adding an identity matrix with a random, again exponentially distributed with rate 10, scalar to each of the covariance matrices. After that, we will calculate the estimated sample covariance using the geodesic path length distance. We again use the `pdMean` function and take the real part of the estimated matrix.

Finally, we estimate it using what will be referred to as 'Geodesic' mean. This is the Frechét mean of the covariances using the geodesic path length distance function on the trace-class extended Hilbert-Schmidt space. This differs from the Thompson covariance estimate, because the matrix that is added to each of the covariances, is not an identity multiplied by a scalar, but a trace-class matrix multiplied by a scalar. To calculate the mean, we use the same function as for the Thompson mean.

With both the Thompson and Geodesic means, there were some issues on the computational side. Many of the covariance matrices are 'barely' positive definite, which is a requirement for the distance functions. This problem is solved by adding a scalar matrix with $\frac{1}{100}$ on the diagonal. Of course, this makes it so that the Geodesic metric doesn't apply to trace-class matrices in practice. Sadly, this is a concession that had to be made for the purposes of computation.

4.2.2 Test statistics

A test statistic is a numerical value calculated from sample data in a statistical hypothesis test. It is used to determine the likelihood of accepting or rejecting a null hypothesis. The test statistic summarizes the information from the sample and allows for the comparison of observed data with what would be expected under the null hypothesis.

In a more common parametric setting, there is often extensive knowledge on the asymptotic behavior of test statistics. When such information is available, it is possible to use the central

limit theorem to determine the p-value of a test based directly on the value of a test statistic, by comparing it to the quantile of a known distribution such as the chi-squared distribution.

For this power test, we will be using a permutation test to determine the p-values. In the context of a permutation test, the test statistic will be calculated from the two samples of the data. Then, by selecting random permutations of the two samples of data and calculating their test statistics, an estimate of the distribution of the test statistic under the null hypothesis is constructed. Then the original test statistic is compared to the constructed distribution and the quantile of the location of the test statistic is the p-value from the test.

For the test statistic, we have four different choices. We will compare the Frobenius distance (section 2.4.1), Procrustes size-and-shape metric (section 2.4.3) and log-Euclidean distance (section 2.4.2).

We calculate all four test statistics during every permutation, so that we can directly compare the results they generate on the exact same data.

4.3 Results

First, we performed the power test using the standard sample covariance as our estimate of the actual covariance. Then we performed permutation tests based on the Frobenius distance, Procrustes shape and size distance and Riemannian, or geodesic, distance between the two sample covariances. This gave us the following results for power:

It can be seen that the Riemannian distance does not work, likely because the matrices are required to be positive definite and the covariances are very close to not being positive definite. So much so, that any small machine calculation errors can make the matrices lose that status. The Procrustes distance seems very good at evaluating the difference between the two sample covariances, while the Frobenius distance shows an expected increase in power for the phonemes that are easier to tell apart. It is likely that there is also an issue with the Procrustes distance, that causes all powers to be 1. To see if these test statistics function at all, it is important to look at the false positive rate. That will tell us if these statistics work as expected, or if they fail as a statistic and have a high false positive rate as well.

4.3.1 False positive rate

Tests with great power are of course very nice to have. But it is important to keep the false positive rate in line with the expectation that a level of significance of 0.05 should provide. What is important to check for, is that the false positive rate should be in the ballpark of 0.05. With only 250 repeats of the permutation test, it can happen that there are some deviations from the expected false positive rate. It would not be unreasonable to expect values in the range of 0.03-0.07 for the false positive rate with this small number of repeats. What we are mostly looking for is that none of the types of mean or distance function consistently show a higher false positive rate.

As can be seen in table 2, the false positive rate for all types of mean and test statistics

Table 1: The power of the permutation test. Calculated from 250 repeats of a permutation test with 250 permutations. The rows correspond with the different two sample tests and methods of averaging the individual covariances. The columns have the different distance functions that provide the test statistic. Questions should be raised about the test power of 1.000, especially in the case of the tests comparing `aa` to `ao` and `iy`Ȧ test with a power of 1.000 is not realistic, but it could be caused by the methodology. We set a seed at the beginning of each of the power tests, which leads us to select the same subsets of data each time and also the same permutations are checked, so if this random selection is one with good power each time, it is not unexpected that multiple tests return the power of 1.000. This could only happen if the actual power is already quite high and we would not expect to see this if the real power of the Riemannian and log-Euclidean test statistics would be around the 0.500 or 0.650 that we see from the Frobenius and Procrustes metric.

		Distance function			
	Mean	Frobenius	Riemannian	Procrustes	log-Euclidean
`ao`					
	Sample	0.472	1.000	0.692	0.992
	Extended	0.500	1.000	0.652	1.000
	Geodesic	0.064	0.064	0.060	0.064
	Thompson	0.064	0.064	0.064	0.064
`iy`					
	Sample	0.608	1.000	0.788	1.000
	Extended	0.576	1.000	0.716	0.996
	Geodesic	0.064	0.064	0.060	0.064
	Thompson	0.064	0.064	0.064	0.064
`sh`					
	Sample	1.000	1.000	0.952	1.000
	Extended	1.000	1.000	0.948	1.000
	Geodesic	0.064	0.060	0.060	0.064
	Thompson	0.064	0.064	0.064	0.060

Table 2: The false positive rate of the permutation test. Calculated from 250 repeats of a permutation test with 250 permutations. The rows correspond with the different two sample tests and methods of averaging the individual covariances. The columns have the different distance functions that provide the test statistic.

		Distance function			
	Mean	Frobenius	Riemannian	Procrustes	log-Euclidean
ao					
	Sample	0.068	0.028	0.044	0.044
	Extended	0.048	0.048	0.044	0.044
	Geodesic	0.04	0.064	0.064	0.064
	Thompson	0.04	0.064	0.064	0.064
iy					
	Sample	0.072	0.056	0.032	0.056
	Extended	0.040	0.052	0.040	0.052
	Geodesic	0.04	0.068	0.064	0.068
	Thompson	0.04	0.068	0.064	0.068
sh					
	Sample	0.06	0.032	0.044	0.036
	Extended	0.052	0.048	0.040	0.052
	Geodesic	0.04	0.068	0.064	0.068
	Thompson	0.04	0.068	0.060	0.068

fall into the range that we expected and none seem far away from the expected 0.05. We can see that the different 2nd phonemes show very similar false positive rate to each other. This can be explained by the fact that we use the same seed for the pseudo-random elements of each test and the same subset of `aa` is probably chosen every time, so half of each of the samples is probably the same for each of the three sample comparisons.

4.4 Discussion

There are many interesting elements to discuss among the results of the data analysis. Some of the results are very promising, while others are quite puzzling and even confusing. First, we will compare the four different methods of calculating the estimated sample covariance. Then we will compare the different test statistics in the second part.

4.4.1 Covariance estimation

The first and most obvious thing to notice is the fact that the Thompson and Geodesic estimation methods do not work. It is not just that the power is worse, but the power that is returned is very similar to the false positive rate. What this tells us is that the means calculated using the `pdMean` are essentially random. There is no difference in the power that is returned under the null hypothesis, compared to the alternative. It is unclear at this point where the problem lies. Maybe this method of calculating the mean does not work due to problems on the computational side with positive definiteness, but it could also be a problem with the `pdMean` function or our methodology.

Next is the difference between the Sample and Extended estimates. These differences are very minor, with most of them being smaller than 0.05 power more or less for each of the test statistics. With our number of repeats for both the power test and reputation test, the differences aren't large enough to notice a real difference. Where the differences can be seen, they are not consistent across the different test statistics and 2nd samples. The Frobenius metric has more power with the Extended estimate with `ao` as 2nd phoneme, but less power with the Extended estimate with `iy` as 2nd phoneme. The Procrustes test statistic shows a slightly higher power using the Sample estimator for all three phoneme, but the difference isn't that large overall.

For both the Riemannian and log-Euclidean test statistics, there is no observable difference, since all the powers are equal to 1, or very close.

4.4.2 Test statistics

The differences between the test statistics are interesting to see. We will again begin with the most immediately noticeable results, the extremely high powers of the Riemannian and log-Euclidean test statistics.

Using the Riemannian distance for a test statistic returns a power of 1 for all three two-sample tests with both the Sample and Extended covariance estimates. This value is so high that it does not seem realistic, and it clearly needs to be investigated further to confirm this

result before we can accept it. It is especially difficult to believe a power of 1.000 in the context of the two-sample test between `aa` and `ao`. These were expected to be difficult to tell apart.

Similarly, the log-Euclidean test statistic also results in powers that are not believable. The difference with the Riemannian test statistic is that the two results with powers of 0.992 and 0.996 show that there is not some clear fault that automatically sets the power to 1.000. It is clear that the test is not perfect, and there is some room for a false negative here.

The Frobenius distance behaves very much as expected. For the two-sample test between `aa` and `ao`, we clearly see that it is very difficult to tell these samples apart, with a power of only around 0.500. The power then increases for the comparison to `iy`, to around 0.600, which is still lower than the often-used target of 0.8. For two phonemes that are very different in the case of `aa` and `sh`, we see that the power with the Frobenius test statistic rises to 1.000, which is in line with expectations. If the Frobenius metric is used and the expectation is that the covariances are not *that* different, it is therefore advisable to have a sample size higher than 25 to increase the power.

With the Procrustes size-and-shape metric, we see a similar progression in the power of the test. For the hardest comparison, the power is around 0.65, rising to 0.75 for the easier test with `iy`, and going up to 0.95 for the test with `sh`. This is a better test statistic for the phonemes that are closer than the Frobenius test statistic, but not quite as good for `sh`.

One possible explanation for these results is that the Frobenius test statistic mostly focuses on the different sizes of the covariances and the Riemannian and log-Euclidean mostly on different shapes. The Procrustes size-and-shape metric, as the name suggests, uses both. This could make it more consistent, but less effective when either extreme is of interest. However, we need to explore this hypothesis before we can discuss it further. To analyze the differences between the four different test statistics and two functioning estimation methods further, we will use simulated functional data with a predetermined covariance matrix that will be generated based on the needs of the study.

Chapter 5

Simulation study

As mentioned at the end of chapter 4, we want to use a simulation study to discover what the power of the permutation test with different test statistics is, based on changes in the covariance structure.

Simulation allows us to see how the different means and test statistics react to differences in the data. Changes can be made in the resolution of the data, the hypothesis that is tested and the parameters of the covariance functions. This can be used to test for very specific hypotheses. In this controlled simulation, we have several research questions that we would like to explore.

We want to know if there are differences in power of permutation tests with different ways of calculating the mean covariance and different test statistics. If there are differences in power, does the nature of the difference between the covariance matrices affect these differences? Specifically, since our covariance function is determined by three parameters, do changes in different parameters have a different effect on the power?

Further, we can now also explore the effect of resolution on the power of the permutation test. Does the power increase by increasing the resolution, or is that effect not as large as increasing sample size. Does the result become more consistent with increased resolution. Do certain test statistics and estimation methods perform better or worse with higher resolutions?

5.1 Simulated data

As mentioned in section 2.5, we can generate functional data from a mean element, a covariance function and random element, all defined on a certain support T at d points.

First, we define the support T as the interval $[0, 1]$ as a sequence of equidistant numbers from 0 to 1 with output length $d \geq 2$, since we need at least 0 and 1 in the support. Next, we define the mean element of the function as $\sin(2\pi t)$, for $t \in T$. We choose to add a mean function in the simulated data to generate a bit of realistic uncertainty in the process of having to subtract the mean function from the sample.

Next to be defined is the covariance function. This function is defined as

$$K(s,t) = a \ \exp(-b|s-t|^c), \tag{5.1}$$

$$\mathscr{K} = a \ \exp(-bD^c), \tag{5.2}$$

where a, b and c are parameters that can be chosen to generate. The first form is how to determine the covariance for any two points s and t and the second version is the covariance operator for a distance operator D. The finite dimensional version of D can be considered a matrix where, for an index set $T = \{1, \ldots, n\}$, matrix entry $a_{i,j} = |i - j|$. The three parameters can all be tweaked and show different behavior of the covariance.

Parameter a can be considered something like the variance. When $s = t$, the covariance simplifies to a on the diagonal, which is related to the variance. The effect of increasing a is somewhat akin to increasing variance, as the functions are much further from the mean and their values are more extreme, which can be seen in figure 5a. Similarly

The random element is obtained from generating uncorrelated vectors, where each entry is $N(0, 1)$ distributed. For this simulation study, we take a grid of points as support from the interval $[0, 1]$, with the density of the support determined by the desired dimension d. We then simulate n sample functions with mean function $\sin(2\pi t)$ and covariance function $K(s, t) = k \ \exp(-c|s - t|^{\mu})$. $K(s, t)$ can be viewed as the $K_{s,t}$ for the covariance matrix K. As the resolution increases, the covariance function approaches the covariance operator. The covariance function is governed by three parameters. The first is k, which is essentially the variance (see $s = t$, the diagonal entries, which reduce to k). Second is c, which introduces exponential scaling of the covariance. Finally, μ introduces a non-linear relation in the covariance based on the distance between s and t, this needs to be smaller than 2.

Given the mean, covariance function K and uncorrelated random normal elements z_i, the data is generated as follows. First, the Cholesky decomposition of covariance matrix K is taken and then multiplied by the random uncorrelated normal vector z_i for sample i. This gives us the sample function i, given by $F_i = \mu + \text{chol}(K)z_i$.

To illustrate the impact of the three parameters of the covariance function, we will simulate some data. All simulation plots show 10 samples simulated at 100 points on the interval $[0, 1]$, with the baseline covariance parameters all set to 1. The first simulation shows all parameters at 1. The second plot has k, the variance, turned up to 10, which clearly increases the spread, looking at the y-axis. The functions are much further apart.

5.1.1 Vector parameters

One issue that we found with the current setup for the covariance function, is that each of the three parameters affects the whole covariance. When we look at the heatmaps of the sample covariances for `aa`, `iy` and `sh` on the same scale, we see that the differences aren't uniform across the matrix, but concentrated in specific areas. This can be seen in figure 3. It could be that some of the test statistics work better on less uniform covariance matrices, which is something worth testing.

To mimic this kind of result, we will also run simulations using two vector parameters. The first of these is what we will call the 'symmetric' vector parameter, and the other is the 'asymmetric' vector parameter. Both the symmetric and asymmetric vector parameters build up the values from the baseline of 1 up to 10. For the symmetric parameter, we replace the first half of the vector, entries 1 to 75, with a sequence from 1 to 25 in value and a sequence from 10

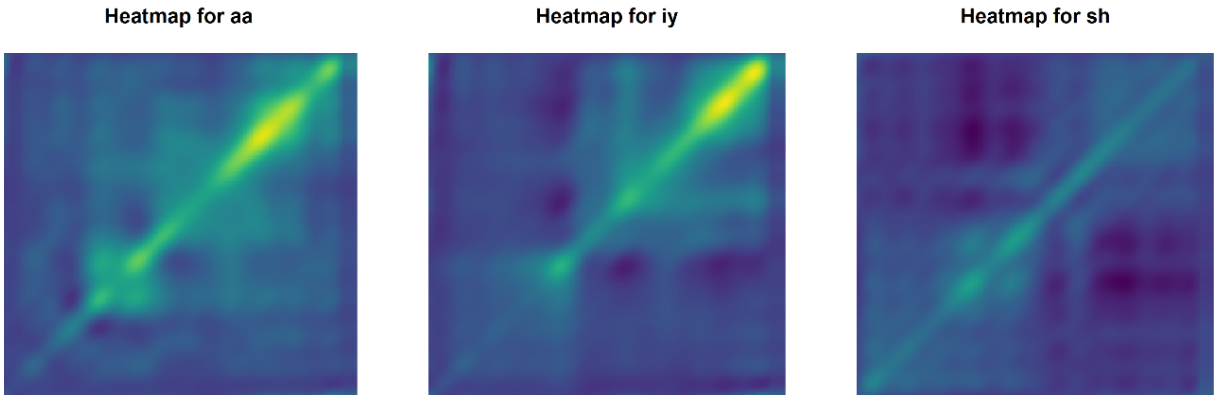


Figure 3: Heatmap of the covariance matrices for `aa`, `iy` and `sh`, all on the same scale for the heatmap. The differences in their covariance structures are not uniform across the matrix, but concentrated in specific areas.

to 1 for the entries 76 to 150. With the asymmetric vector, we do roughly the same, however we have the increasing sequence for entries 1 through 30 and the decrease from 31 through 60 and leave 61 through 150 at the baseline value of 1. The functions used for the parameter vectors can be seen in figure 4.

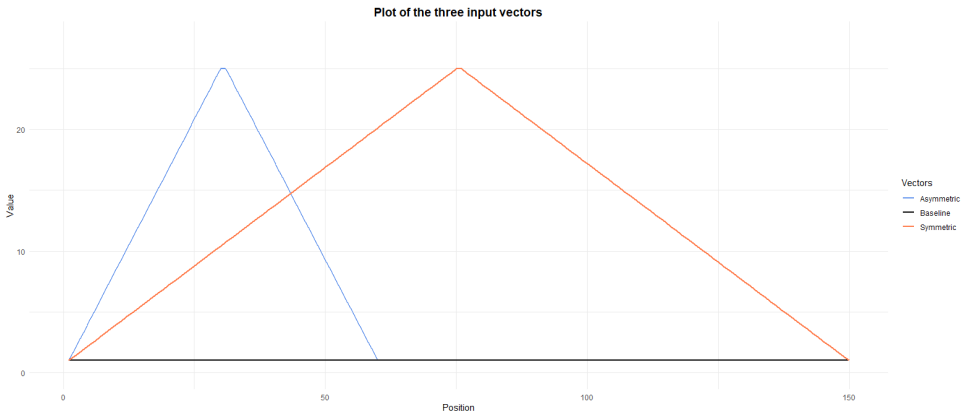
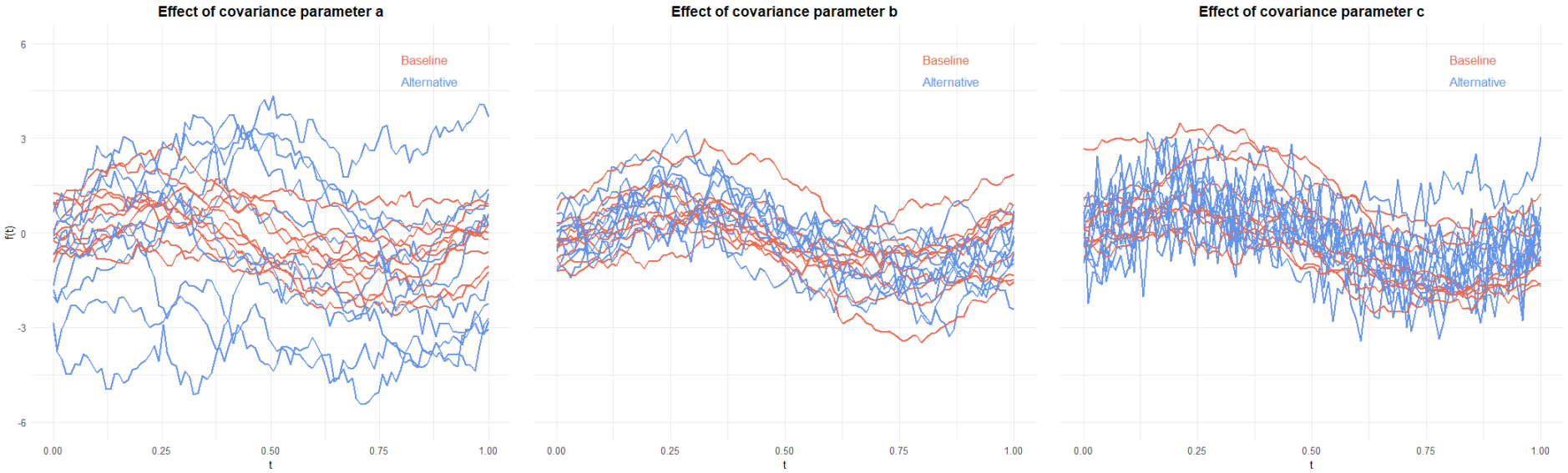


Figure 4: Function showing the baseline value and two vector parameters used..

With these vector parameters, we can again simulate the data and perform our power test. To show the differences in structure between the baseline and vector parameters, we can plot the same heatmaps again for some of the new combinations.



(a) Comparison of functions with baseline covariance (1, 1, 1) with alternative covariance (5, 1, 1).

(b) Comparison of functions with baseline covariance (1, 1, 1) with alternative covariance (1, 5, 1).

(c) Comparison of functions with baseline covariance (1, 1, 1) with alternative covariance (1, 1, $\frac{1}{5}$).

Figure 5: The effect of a change in the three covariance parameters on the simulated functions. For each of the three images, we plot 10 randomly simulated functions based on a mean function of $\sin(2\pi t)$ and the covariance with the parameters listed below. The definition of the covariance function and its parameters a, b and c is as in (5.1). A change in parameter a causes a big increase in what could be considered the variance of the functions, since they are much further apart. A change in parameter b increases the sharpness of the covariance quite a bit. Changes in parameter c cause sharp changes between two points that could be very close together. This is almost akin to noise and could be mistaken as such.

5.2 Method

At the start of the simulation run, the seed for the pseudo-random elements is set at 19231031, and we set the parameters for the covariance function to those required. Two populations are then fully simulated, both the base population using the covariance parameters $(1, 1, 1)$ and the alternative. For both populations, we simulate 400 functions, to have the same total population size as that of the phoneme data. The resolution of the functions is also chosen at the start, so only the function values for the required number of points (10, 50, 150 or 250) are simulated.

After simulating both populations in this way, the rest of the methodology that is used is exactly the same as it was for the data analysis on the phoneme data, which can be found in section 4.2.

5.3 Results

In the results, we see very reasonable values for the power of each of the simulation power tests overall.

One noticeable thing is that for changes in parameters a and b, the power of the test using the Riemannian and log-Euclidean test statistics vanish to zero. They start at values that could be considered as a false positive, but they vanish to zero with higher values. This behavior is seen for both the Sample and Extended covariance estimators. It is not consistent between the different resolutions, however. While the power does always vanish when the difference is based on parameter a, it does not vanish for changes in parameter b from the resolution of 50 or higher.

Further, we can see that the power of all four test statistics and for both of the estimation methods increases with higher resolution. An increase in resolution does therefore clearly improve the power of the test.

The Procrustes and Frobenius test statistics come out better in the simulation compared to the data analysis. Especially, the Frobenius metric seems very powerful at smaller or medium-sized differences in the covariance parameters.

For the full results of the results of the simulation power test with the Sample covariance estimator, look to table 6. The results from the test using the Extended estimator were extremely similar and can be found in 4. We have also checked the results for the Thompson estimator, and they are all, similarly to the data analysis results, at the level of the false positive rate. For the full results of the Thompson metric, look at appendix A. Since the results for the Thompson and Geodesic estimation methods were the same, we didn't check the latter seeing as the results were all around the level of significance.

Next, we look at the results for the simulations using a vector instead of scalar as parameter. For these tests, we only use the Sample covariance estimator, since the results from different estimations are all quite similar so far. Using only one vector doesn't really give any of the tests good power, except for the Frobenius distance. The results using only one vector parameter are not dissimilar to the results using a 2 as scalar for one of the parameters, except for parameter c. Overall, these results show very low power.

Replacing two of the scalars with a vector paints a very different picture. With this test, all the powers are extremely high. We don't see quite as many differences between the different tests, but a few things stick out. First, we see that it is harder to get high power for the asymmetric vectors, most likely due to the smaller overall sum of values in the vectors. We can also clearly see that the Riemannian and log-Euclidean distance functions work the best with the third parameter, with their powers significantly lower when only parameters a and b are changed. Finally, using the asymmetric vectors for parameters b and c, we can finally see a situation where the Riemannian and log-Euclidean distances are clearly better.

5.4 Discussion

Looking at the simulation results using scalar parameters, it is very clear that the Riemannian and log-Euclidean test statistics give very similar results. This also coincides with the results from the data analysis. On the other side, the Frobenius metric gives the most clear result when the variance of the functions around the mean is the highest, as is the case when changing parameter a. It also performs worse than the other test statistics on parameter c, which has more to do with the shape of the covariance. The test statistic based on the Procrustes metric seems like the best compromise, showing good results on all the parameters, but never being the outright best parameter.

These results communicate to us that there is no clear 'best' test statistic between the four choices. Which test statistic to choose in the context of hypothesis testing does become quite a difficult one, based on these simulation results. With the knowledge of what type of differences in covariance you are testing for, there are clear right, but also very wrong choices. When the difference is like a change in parameter a, then the choice of Riemannian or log-Euclidean statistic, makes the difference undetectable.

The results from the simulated data again confirm that our methodology of applying a Frechét mean with either the Thompson or Geodesic metric does not work. There is also no real difference between using the Sample or Extended covariance estimator.

The choice of scalar parameters is an easy one in the context of simulation. However, a scalar parameter affects the whole covariance matrix evenly. This is not often the case in reality, as can be seen in the heatmaps of the covariance matrices of the phoneme (figure 3). To hopefully get somewhat closer to this type of more complicated covariance structure, we applied the vector parameters, both symmetric and asymmetric variants.

These results show more differences in the behavior of the different test statistics. Once again, the Frobenius distance is often the most effective. However, as seen on the final row of table 5, we do find evidence in simulation that the Riemannian and log-Euclidean distance functions can be much more powerful test statistics compared to the Frobenius and Procrustes metric. This shows that the results from the power test on the phoneme data can almost be reproduced and that the concept of some test statistics having a power near 1.000, with others much lower, is not out of the question.

There are extremely high differences in the effectiveness of test statistics depending on the type of difference between two covariance structures. This could make it very hard for researches to choose the correct test statistic to use when they want to find a difference between two samples based on covariance.

Table 3: Power of the different test statistics on the simulated data. All use the Sample covariance as an estimation method. On the left-hand side, three categories are distinguished by resolution, that is the number of points where the function is measured are set to 10, 50 or 250. Then next to it are the parameters for the covariance function, (a, b, c), as used in function (5.1).

		Distance function			
	Parameters	Frobenius	Riemannian	Procrustes	log-Euclidean
10					
	(2, 1, 1)	0.388	0.048	0.100	0.060
	(5, 1, 1)	0.980	0.000	0.336	0.012
	(25, 1, 1)	1.000	0.000	0.948	0.000
	(1, 2, 1)	0.280	0.056	0.112	0.056
	(1, 5, 1)	0.884	0.016	0.272	0.044
	(1, 25, 1)	1.000	0.000	0.632	0.012
	(1, 1, $\frac{1}{2}$)	0.192	0.084	0.136	0.088
	(1, 1, $\frac{1}{5}$)	0.420	0.088	0.212	0.104
	(1, 1, $\frac{1}{25}$)	0.568	0.064	0.272	0.100
50					
	(2, 1, 1)	0.648	0.040	0.196	0.032
	(5, 1, 1)	1.000	0.000	0.544	0.004
	(25, 1, 1)	1.000	0.000	0.972	0.000
	(1, 2, 1)	0.220	0.132	0.172	0.112
	(1, 5, 1)	0.620	0.268	0.348	0.236
	(1, 25, 1)	0.996	0.840	0.868	0.792
	(1, 1, $\frac{1}{2}$)	0.580	0.196	0.692	0.440
	(1, 1, $\frac{1}{5}$)	0.916	0.456	0.652	0.856
	(1, 1, $\frac{1}{25}$)	0.972	0.660	0.060	0.932
250					
	(2, 1, 1)	0.100	0.164	0.132	0.140
	(5, 1, 1)	0.820	0.048	0.300	0.060
	(25, 1, 1)	1.000	0.000	0.868	0.000
	(1, 2, 1)	0.712	0.108	0.184	0.092
	(1, 5, 1)	0.988	0.104	0.436	0.088
	(1, 25, 1)	1.00	0.648	0.920	0.632
	(1, 1, $\frac{1}{2}$)	0.532	0.480	0.264	0.392
	(1, 1, $\frac{1}{5}$)	0.720	0.936	0.540	0.848
	(1, 1, $\frac{1}{25}$)	0.784	0.984	0.716	0.968

Table 4: Power of the different test statistics on the simulated data. All use the Extended estimation method, using the canonical metric on the extended HS-space. On the left-hand side, three categories are distinguished by resolution, that is the number of points where the function is measured are set to 10, 50 or 250. Then next to it are the parameters for the covariance function, (a, b, c), as used in function (5.1).

		Distance function			
	Parameters	Frobenius	Riemannian	Procrustes	log-Euclidean
10					
	(2, 1, 1)	0.532	0.048	0.108	0.320
	(5, 1, 1)	0.984	0.004	0.340	0.012
	(25, 1, 1)	1.000	0.000	0.912	0.000
	(1, 2, 1)	0.268	0.052	0.116	0.044
	(1, 5, 1)	0.876	0.032	0.280	0.040
	(1, 25, 1)	1.000	0.012	0.676	0.036
	(1, 1, $\frac{1}{2}$)	0.176	0.116	0.148	0.096
	(1, 1, $\frac{1}{5}$)	0.356	0.136	0.228	0.152
	(1, 1, $\frac{1}{25}$)	0.508	0.132	0.264	0.140
50					
	(2, 1, 1)	0.620	0.028	0.212	0.056
	(5, 1, 1)	1.000	0.000	0.552	0.004
	(25, 1, 1)	1.000	0.000	0.980	0.000
	(1, 2, 1)	0.216	0.156	0.180	0.128
	(1, 5, 1)	0.600	0.332	0.404	0.296
	(1, 25, 1)	1.000	0.868	0.916	0.852
	(1, 1, $\frac{1}{2}$)	0.096	0.576	0.244	0.484
	(1, 1, $\frac{1}{5}$)	0.248	0.948	0.468	0.864
	(1, 1, $\frac{1}{25}$)	0.492	0.980	0.704	0.960
250					
	(2, 1, 1)	0.092	0.148	0.100	0.120
	(5, 1, 1)	0.788	0.048	0.268	0.060
	(25, 1, 1)	1.000	0.000	0.872	0.000
	(1, 2, 1)	0.648	0.084	0.148	0.068
	(1, 5, 1)	0.980	0.088	0.388	0.100
	(1, 25, 1)	1.00	0.620	0.948	0.600
	(1, 1, $\frac{1}{2}$)	0.524	0.484	0.224	0.364
	(1, 1, $\frac{1}{5}$)	0.664	0.952	0.560	0.880
	(1, 1, $\frac{1}{25}$)	0.752	0.984	0.716	0.972

Table 5: Power of the different test statistics on the simulated data. These all use vectors for the second sample instead of different scalars as parameters. The covariance is estimated using the Sample covariance estimator.

	Distance function			
Parameters	Frobenius	Riemannian	Procrustes	log-Euclidean
(sym, 1, 1)	1.000	0.008	0.688	0.016
(1, sym, 1)	0.440	0.520	0.328	0.384
(1, 1, $\frac{1}{sym}$)	0.272	0.952	0.488	0.892
(asym, 1, 1)	0.996	0.000	0.572	0.004
(1, asym, 1)	0.500	0.260	0.220	0.196
(1, 1, $\frac{1}{asym}$)	0.440	0.588	0.244	0.436
(sym, sym, 1)	1.000	0.384	0.984	0.628
(sym, 1, $\frac{1}{sym}$)	1.000	0.996	0.996	0.996
(1, sym, $\frac{1}{sym}$)	0.996	0.992	0.976	0.988
(asym, asym, 1)	0.996	0.124	0.720	0.180
(asym, 1, $\frac{1}{asym}$)	0.988	0.924	0.856	0.896
(1, asym, $\frac{1}{asym}$)	0.464	0.864	0.436	0.756

Chapter 6

Conclusion

Functional data analysis is an area of statistics that is increasing in relevance as our techniques of capturing data increase in resolution. Most of the knowledge that currently exists on functional data is focused on comparing the mean elements of samples of data, but much of the information in any given dataset is also contained within the covariance structure. To improve our knowledge around functional data, it is imperative to improve our understanding and range of methods of analyzing functional data, both from a mathematical and a practical perspective. This will be especially important if the dimensionality of this, inherently infinite dimensional, type of data continues to be captured with higher resolutions.

As mentioned earlier, many methods that are used for the analysis of functional data rely on using a finite dimensional representation of functional data, such as those using a spectral decomposition (2.30) or the Karhunen-Loève theorem (2.25). Other methods make direct use of the fact that the way we measure, store and use functional data is finite dimensional in nature. Amongst these are the methods that make use of the Extended Hilbert-Schmidt space, such as the Thompson metric, explained in section 3.1.

We wanted to expand on that second method of expanding the Hilbert-Schmidt space that contains the covariance operators. Our idea was to expand this space not with scalar operators, as was the case in the work of Lawson and Lim (2013b), but with trace-class operators. This ensured that the resulting space still had the trace-class property. Any norms and their resulting distance functions defined on this space could also be used in the context of a Frechét mean to calculate an estimate for the mean covariance of a population of functions.

These lofty goals on the mathematical side didn't work as planned, however. Defining the space was difficult, since elements in the trace-class extended Hilbert-Schmidt space are unique. This led to further issues, where the inner product could not be defined in a way that it is unique, meaning that there was no inner product defined on the space. We did still explore what the inner product would look like if a way was found to define the elements in the space as unique, and how the inner product would give us a canonical metric on the space. Further, we also used the properties of the space to define a metric using the path length of the geodesic between two covariances. In the discussion, we will look at potential paths that could resolve the problem of uniqueness of elements in the trace-class extended Hilbert-Schmidt space.

Further, we also found that many computational issues arise when using a trace-class extension in practice. Covariances are positive definite by definition and therefore also in practice, but they are not *very* positive definite. Often, they are only a few rounding errors in `R` away from not being positive definite anymore. Using a scalar extension, with the scalar set to more than a certain value (in our experience often around 10^{-2}), made the covariances more positive and helped to avoid any problems with matrices not being positive definite in practice. Using a trace-class extension didn't help with the positivity of matrices and led to problems with computation in many more scenarios.

After our the new trace-class extended Hilbert-Schmidt space was not well-defined from a mathematical perspective, it was attempted to implement it in practice to see if it does work on an actual dataset. We chose to compare four different methods of estimating the sample covariance and four different test statistics in a power test experiment, where our p-values were calculated using a permutation test. This power test was performed for three different two-sample tests from the `phoneme` data that is available in the `fda` package in `R`. We expect that the three different two-sample tests show a nice progression in power from the most difficult comparison to the easiest.

From these power tests, we found that the results were not as straightforward as hypothesized. One big observation was that the implementation of the Thompson and Geodesic estimation methods did not work. Further, the Riemannian test statistic, which is the same as the geodesic path length metric, and the log-Euclidean statistic have extremely high power in all three two-sample tests. A test with a power close to one is hard to believe, and therefore something that requires to be checked, both by looking at the false positive rate of the test and by reproducing it in simulations. Our hypotheses were met by the power of the Frobenius and Procrustes test statistics, these did show the nice progression that we were expecting. The false positive rate of all the tests were very close to our level of significance, so all seemed valid from that perspective.

To both validate the results from the data analysis and further study the behavior of the different test statistics, we also performed a simulation study. In this part, the covariance functions were defined with three tunable parameters. These covariance functions were then combined with a random element and mean function to simulate the data. The methodology used for the data analysis was also applied to analyze the simulation data.

In the results, we could see a nice progression of power of the tests with both increasing parameter value and increasing resolution. Higher resolutions particularly seemed to be better for both the Riemannian and log-Euclidean test statistics. On the parameter side, we saw that changes in parameter a caused high power in the tests using Frobenius and Procrustes test statistics, while the power of the test with Riemannian and log-Euclidean test statistics went to 0. Parameter b seemed quite balanced and all test statistics were able to detect the changes, with the Frobenius statistic still quite good. For testing differences in parameter c, you would want to use the Riemannian or log-Euclidean test statistics for the highest power.

Finally, the simulation also revealed that there is very little difference between the different estimators for the mean. It also confirmed that the Thompson metric and therefore also the Geodesic metric, which uses the same function to calculate the mean, don't work with our methodology. Testing differences between different scalar parameter sets didn't show us similar results to the data analysis yet, with the Riemannian and log-Euclidean test statistics clearly being better.

The last simulations that were performed were those with vectors instead of scalar parameters. This was meant to provide some non-uniformity to the covariance matrices. When looking at heatmaps of the covariances of the phoneme data, those weren't uniform along the diagonal, but those simulated with scalars were. To address this, we introduced a symmetric and asymmetric vector to try and mimic that structure.

After running the power test with the vector parameters, we found much clearer results than the original simulation. Parameter a can still clearly only be measured by the Frobenius and Procrustes metric, while parameter b is still easy to find by all and changes in parameter c as a vector are much easier to recognize with the Riemannian and log-Euclidean test statistics. We have also finally found some evidence that the results of the data analysis, with extremely high powers for the Riemannian and log-Euclidean test statistics, are valid. Both the vectors $(1, 1, \frac{1}{sym})$ and $(1, asym, \frac{1}{asym})$ show a large discrepancy in the powers of the Frobenius and Riemannian/log-Euclidean test statistics. This can indicate that with some more tweaking, the results of the data analysis could be matched by simulated data.

Chapter 7

Discussion

In this discussion section we will talk about some of the limitations of the methodology and problems encountered in this thesis. Where possible, I will put thought into solutions or at least attempt to point in the direction of solutions.

7.1 Uniqueness of elements

One of the first issues that was encountered was the problem with uniqueness of elements in the trace-class extended Hilbert-Schmidt space. This was also mirrored in the problems with defining the inner product, which also was ill-defined because it was impossible to recover the split between the covariance operator and trace-class element in the way it was done for the inner product for the extended Hilbert-Schmidt space as in equation (3.2). While this problem hasn't been solved in this thesis, several directions for solutions were thought of.

One of the possible solutions to the uniqueness problem is some sort of conditioning on the added trace-class element. We mostly considered diagonal trace-class elements in this thesis. A possible solution could be some condition on the trace-class element that makes it distinguishable and unique from the covariance. This is a similar line of thinking to how the scalar element is unique with respect to the covariance operator and can easily be recovered. It could be hard to find such an element without violating the requirement that each element of the trace-class Hilbert-Schmidt space has the same properties as covariance operators, which was our starting point.

For the problem with defining the inner product, there could be an easier solution. After extending each operator in the Hilbert-Schmidt space with a diagonal trace-class element, we need to define the inner product in a way that it works similarly to the one defined in equation (3.2). Our idea was to use a decomposition on the extended Hilbert-Schmidt space, that is unique for each element. If we extended the space with an operator Σ, then the decomposition of $A \in \mathscr{P}$ as:

$$A = \bar{A} + \lambda\Sigma, \quad \lambda := \sup\{\ \lambda \mid A - \lambda\Sigma \text{ is positive definite }\}, \tag{7.1}$$

where $\bar{A} = A - \lambda\Sigma$. Using this decomposition, we could then define the inner product

$$\langle A, B\rangle_{\mathscr{P}} := \langle \bar{A}, \bar{B}\rangle_{HS} + \lambda\mu\|\Sigma\|_F^2. \tag{7.2}$$

This is one possibility that could solve the issue of the inner product not being well-defined. However, this does nothing for the issue with uniqueness in the trace-class extended Hilbert-Schmidt space. A solution to that problem would most likely also solve the problem with the definition of the inner product.

7.2 Computational constraints

Positive definite matrices have been the subject of a lot of mathematical research over the last years (Bhatia, 2009). These objects are not only important in statistical analysis of covariance matrices, but also in convex and semidefinite programming problems and as kernels in machine learning problems (Lawson & Lim, 2013b). When dealing with such matrices, positive definiteness can be an issue in computational scenarios due to problems wiht rounding.

This is specifically the case when matrices have many very small values or zeroes. These issues come up most often when dealing with sparse or trace-class positive definite matrices. There were two different issues that came up in our methodology. The first was issues with positive definiteness for basic functions that deal with matrices, such as taking a square root or logarithm. The second issue was with the functions from the `pdSpecEst` package, which also rely on positive definiteness.

For the first category, we have dealt with it by taking the eigenvalue decomposition of the matrix and setting all eigenvalues to the machine tolerance level, which is a very small positive value. We then calculate the matrix square root on the new, slightly positive eigenvalues and then reproduce the output matrix from the product of the eigenvalues and eigenvectors.

The second category of functions needs a matrix as an input and cannot be applied just to the eigenvalues. Recalculating the matrices with all eigenvalues set to at least the machine tolerance level did not work for these functions, it was still not positive enough. To combat this issue of positive definiteness, we added a scalar matrix with $\frac{1}{100}$ as scalar to each of the covariances. This was, after trying manually, the smallest scalar that didn't occasionally cause issues with positive definiteness.

These solutions are clearly imperfect. While the values added are quite small, they could make a difference to the results. Especially adding the scalar matrix is a solution with a big impact, since it is basically like using the extended Hilbert-Schmidt space, losing the trace-class property.

To improve on this study, using better methods to deal with these rounding errors is imperative. Potential avenues of solution can be found in Chang et al. (1997) or Scott and Davis (2006). For future applications of similar methods, we recommend looking into more established solutions, such as those mentioned.

7.3 Estimation methods

We wanted to compare four estimation methods to determine the mean covariance of a sample. These are all the Frechét mean of the covariances with different distance functions. For the

first two somewhat similar estimation methods, we use the Frobenius distance function and the canonical distance on the extended Hilbert-Schmidt space. The other two estimation methods depend on the geodesic path length distance. The Thompson estimator being applied to covariances that have had a scalar matrix with random scalar added, and the Geodesic metric to covariances with an added trace-class element with random scalar.

The first two estimation methods are very similar. With some of the tests, there are some minor differences in power, but the difference never exceeded 0.05 by much. This shows that the choice between these two estimation methods does not seem that relevant. To see if one is slightly better than the other, it could be interesting to repeat the study with a higher number of permutations and repeats for the power test.

With the Thompson and Geodesic estimation methods, there seems to be an issue. Theoretically, these are distance functions that should work in a Frechét mean. In practice with our methodology, they don't work. All the powers came in around the level of significance, so it seems like the estimated means are essentially random. We have not found any reason for this behavior, so this would need to be looked into more extensively, if these estimation methods are to be used.

One other constraint with our methods are that they rely on the assumption that the functional data follows a normal distribution. For the estimation of covariance from non-Gaussian functional data, other methods need to be applied. An additional issue with non-Gaussian functional data is that it is not as easy to identify the distribution of the data in comparison to regular parametric statistical settings. Kraus and Panaretos (2012) explore the idea of resistant second order functional data analysis for the situations where the distribution is not surely Gaussian. Other work in the development of robust methods that aren't as susceptible to outliers in high dimensional cases, such as functional data, can be found in Raymaekers and Rousseeuw (2019).

7.4 Power of permutation test

Our hypothesis for the power of the permutation test in the data analysis was that we would be seeing a relatively low power for the two sample test with `aa` and `ao`, and see that power increase for the test with `aa` and `iy`, and further increase for `aa` and `sh`. This hypothesis was correct for both the Frobenius and Procrustes test statistics.

However, with the Riemannian and log-Euclidean test statistics, the power of the test was 1.000, or close to it. A power of 1 is not something realistic in statistics, especially when there was an expectation of lower power for the test with `aa` and `ao`. There may be several reasons why the power is this high for these test statistics.

First off, the power is most likely equal to 1.000, because of the combination of a low amount of permutations at only 250 and a small amount of repeats of the power test at only 250. In reality, the power may be slightly below 1.000, but because of the low amount of repeats, there is more variance in the result of such a test and with the chosen seed we could be in the upper

ranges of the variance. Repeating the experiment with more permutations and such could give us a better idea of what the power of the test truly is.

Secondly, our hypothesis could simply be incorrect for these test statistics. The covariances of the samples were close enough in the aspects that are tested for by the Frobenius and Procrustes test statistics to return values confirming our hypothesis. However, it could be that there are other differences in the covariance that are quite large and only, or mostly, picked up by the Riemannian and log-Euclidean test statistics.

This second possibility was something that we wanted to test for in a simulation study. Could a situation be recreated, where the Riemannian and log-Euclidean test statistics have a very high power, while the Frobenius and Procrustes test statistics have low power? This situation was able to be recreated, both by using the symmetric vector for parameter c, as well as using asymmetric vectors for parameters b and c. The powers for the Riemannian and log-Euclidean test statistics weren't 1.000 in these situations, but the results were not far off. We also managed to find the opposite result in almost any test where parameter a was changed. This vastly increased the power for the Frobenius and Procrustes tests, but also reduced the power of the other two to 0.

Next steps in this line of research would be to find a set of parameters for a covariance matrix used in a simulation that approaches the phoneme. That way more research can be done into the specific scenario. Next, it could be interesting to find data sets that can be compared in a two-sample test where the Frobenius and Procrustes metrics are more effective, to show that the situation where the difference is dominated by simulation parameter a can also be found in actual data.

7.5 Which test statistic to choose

One of the questions that naturally gets asked in any study regarding the power of statistical tests, is which test is the best to use for a two-sample test such as the ones performed in this study. The answer to this question is extremely difficult in this situation, because of the vast differences in power based on the changes of different parameters in the simulation setting. There are no easy answers to this question, but there may be ways to approach an answer.

One such way is through further simulation studies. If it is possible to simulate the whole space of covariance matrices based on a certain set of parameters, these can be changed and the effect of each of those changes can be observed. With increasing knowledge about the role of these parameters in the covariances, experts could hypothesize about the type of difference in covariance structure and choose the test statistic appropriately. In our simplified version with three parameters, and expert could decide on the Frobenius statistic if they expect the difference to be in parameters a of b and the Riemannian test statistic if they expect the difference to be with respect to parameter c.

The second method that could be explored is developing a methodology that calculates different test statistics and where the outcome is based on a combination of these different

statistics. In an attempt at such a method, it is very important to check the false positive rate and make sure that remains at the desired level of significance.

Bibliography

Arsigny, V., Fillard, P., Pennec, X., & Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Analysis Applications*, *29*, 328–347. https://doi.org/10.1137/050637996

Bhatia, R. (2009). *Positive definite matrices*. Princeton University Press. https://doi.org/10.1515/9781400835488

Brown, P. J. (2020). *Time series forecasting with python: An introduction to time series modeling and forecasting techniques*. Springer.

Cai, T. T., & Hsing, R. J. (2015). *Theoretical foundations of functional data analysis with an introduction to linear operators*. John Wiley & Sons.

Chang, X.-W., Bunch, J. R., & Larsen, R. W. (1997). Handling the positive definiteness constraint in rounding error prone situations. *SIAM Journal on Matrix Analysis and Applications*, *18*(3), 583–597. https://doi.org/10.1137/S0895479895296286

Cole, R. (2014). Live-cell imaging: The cell's perspective. *Cell Adhesion & Migration*, *8*(5), 452–459. https://doi.org/10.4161/cam.28348

Coppersmith, D., & Winograd, S. (1990). Matrix multiplication via arithmetic progressions. *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, 1–6.

Dryden, I. L., Koloydenko, A., & Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, *3*(3), 1102–1123. https://doi.org/10.1214/09-aoas249

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Handwriting recognition with deep recurrent neural networks. *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, 6645–6649.

Hackmann, T. (2023). Thesis project toby hackmann. https://github.com/toby-hackmann/thesis-fda

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, *23*(1), 73–102. Retrieved July 5, 2023, from http://www.jstor.org/stable/2242400

Hlávka, Z., Hlubinka, D., & Koňasová, K. (2022). Functional anova based on empirical characteristic functionals. *Journal of Multivariate Analysis*, *189*, 104878. https://doi.org/https://doi.org/10.1016/j.jmva.2021.104878

James, G., & Titterington, D. (2005). Functional linear discriminant analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 191–221. https://doi.org/10.1111/j.1467-9868.2005.00495.x

Jiménez Gamero, M. D., & Franco Pereira, A. (2021). Testing the equality of a large number of means of functional data. *Journal of Multivariate Analysis*, *185*, 104778. https://doi.org/10.1016/j.jmva.2021.104778

Karhunen, K. (1947). Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae. Series A I. Mathematica*, *37*, 1–79. https://doi.org/10.5186/aasfm.1947.3701

Kashlak, A. B., Aston, J. A. D., & Nickl, R. (2018). Inference on covariance operators via concentration inequalities: K-sample tests, classification, and clustering via rademacher complexities. *Sankhya A*, *81*(1), 214–243. https://doi.org/10.1007/s13171-018-0143-9

Kraus, D., & Panaretos, V. M. (2012). Dispersion operators and resistant second-order functional data analysis. *Biometrika*, *99*(4), 813–832. Retrieved August 5, 2023, from http://www.jstor.org/stable/41720736

Larotonda, G. A. (2008). Nonpositive curvature: A geometrical approach to hilbert-schmidt operators. *Acta Applicandae Mathematicae*, *102*(1), 85–102. https://doi.org/10.1007/s10440-008-9249-5

Lawson, J., & Lim, Y. (2013a). The least squares mean of positive hilbert–schmidt operators. *Journal of Mathematical Analysis and Applications*, *403*(2), 365–375. https://doi.org/10.1016/j.jmaa.2013.02.013

Lawson, J., & Lim, Y. (2013b). Weighted means and karcher equations of positive operators. *Proceedings of the National Academy of Sciences*, *110*(39), 15626–15632. https://doi.org/10.1073/pnas.1313640110

Loève, M. (1946). Fonctions aléatoires de second ordre. *Comptes rendus de l'Académie des Sciences*, *222*, 439–441. https://doi.org/10.1016/s1631-073x(46)90103-1

Marquez, Y., Ramsay, J., & Silverman, B. (2008). Functional canonical correlation analysis. *Journal of the American Statistical Association*, *103*(482), 208–219. https://doi.org/10.1198/016214508000000921

Masarotto, V., Panaretos, V. M., & Zemel, Y. (2019). Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhyā: The Indian Journal of Statistics*, *81*(A), 172–213. https://doi.org/https://doi.org/10.1007/s13171-018-0130-1

Masarotto, V., Panaretos, V. M., & Zemel, Y. (2022). Transportation-based functional anova and pca for covariance operators.

Panaretos, V. M., Kraus, D., & Maddocks, J. H. (2010). Second-order comparison of gaussian random functions and the geometry of dna minicircles. *Journal of the American Statistical Association*, *105*(492), 670–682. https://doi.org/10.1198/jasa.2010.tm09075

Petersen, K. B., & Pedersen, M. S. (2008). *The matrix cookbook*. Technical University of Denmark. https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Pigoli, D., Aston, J. A., Dryden, I. L., & Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika*, *101*(2), 409–422. https://doi.org/10.1093/biomet/asu008

R Core Team. (2021). *R: A language and environment for statistical computing* [R version 4.2.1]. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*. Springer. http://www.worldcat.org/isbn/9780387400808

Ramsay, J. O., Hooker, G., & Graves, S. (2020). *fda: Functional data analysis* [R package version 5.1.0]. https://cran.r-project.org/package=fda

Raymaekers, J., & Rousseeuw, P. J. (2019). Fast robust correlation for high-dimensional data. *Technometrics*, *63*(2), 184–198. https://doi.org/10.1080/00401706.2019.1677270

Scott, J. A., & Davis, P. J. (2006). Numerical stability of cholesky factorization methods for large sparse positive definite matrices. *SIAM Journal on Matrix Analysis and Applications*, *28*(4), 1001–1026. https://doi.org/10.1137/040612124

Strang, G. (2009). *Introduction to linear algebra*. Wellesley-Cambridge Press.

Strassen, V. (1969). Gaussian elimination is not optimal. *Numerische Mathematik*, *13*(4), 354–356.

Vu, T., Wrobel, J., Bitler, B., Schenk, E., Jordan, K., & Ghosh, D. (2022). Spf: A spatial and functional data analytic approach to cell imaging data. *PLoS computational biology*, *18*(6), e1009486. https://doi.org/https://doi.org/10.1371/journal.pcbi.1009486

Appendix A Thompson simulation results

Table 6: Power of the different test statistics on the simulated data. All use the Thompson covariance as an estimation method. On the left-hand side, three categories are distinguished by resolution, that is the number of points where the function is measured are set to 10 or 50. Then next to it are the parameters for the covariance function, (a, b, c), as used in function (5.1). Seeing almost no difference between the simulation with 10 and 50 resolution, we decided not to run the simulation with 250 resolution.

		Distance function			
	Parameters	Frobenius	Riemannian	Procrustes	log-Euclidean
10					
	(2, 1, 1)	0.048	0.040	0.040	0.040
	(5, 1, 1)	0.044	0.044	0.040	0.044
	(25, 1, 1)	0.044	0.044	0.040	0.044
	(1, 2, 1)	0.048	0.040	0.040	0.040
	(1, 5, 1)	0.048	0.040	0.040	0.040
	(1, 25, 1)	0.048	0.040	0.040	0.040
	(1, 1, $\frac{1}{2}$)	0.048	0.040	0.040	0.040
	(1, 1, $\frac{1}{5}$)	0.048	0.040	0.040	0.040
	(1, 1, $\frac{1}{25}$)	0.048	0.040	0.040	0.040
50					
	(2, 1, 1)	0.048	0.040	0.052	0.040
	(5, 1, 1)	0.048	0.048	0.052	0.048
	(25, 1, 1)	0.060	0.048	0.060	0.48
	(1, 2, 1)	0.048	0.040	0.044	0.040
	(1, 5, 1)	0.048	0.040	0.044	0.040
	(1, 25, 1)	0.048	0.040	0.044	0.040
	(1, 1, $\frac{1}{2}$)	0.048	0.040	0.044	0.040
	(1, 1, $\frac{1}{5}$)	0.048	0.040	0.044	0.040
	(1, 1, $\frac{1}{25}$)	0.048	0.040	0.044	0.040