Bugni and Horowitz (2021) Permutation Tests for the Equality of Distributions of Functional Data

Master's Thesis presented to the
Department of Economics at the
Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of
Master of Science (M.Sc.)

Supervisor: Prof. Dr. Dominik Liebl

Submitted in July 2022 by:
Jakob R. Juergens
Matriculation Number: 2996491

Contents

1 Introduction

In modern economics, it is becoming increasingly common to use data measured at very high frequencies. As the frequency of observing a variable increases, it often becomes natural to view the data not as sequences of distinct observations but as smooth functions describing a variable. This idea, to think of observations as measurements of a continuous process, is the motivating thought behind functional data analysis, a branch of statistics that has its origins in the 1940s and 1950s in the works of Ulf Grenander and Kari Karhunen. Today, methods from functional data analysis are still relatively exotic in economics, but they are beginning to gain traction. A typical question in economics is whether observations from two or more data sets, e.g., data generated by treatment and control groups, stem from the same data generating process. This question also occurs in functional data analysis, where each observation in a data set is itself a smooth curve.

This thesis presents a test developed by Bugni and Horowitz 2021 trying to answer this question. At the same time, this thesis provides the necessary theoretical background to understand the construction of the test statistic. Therefore, Section 2 introduces the necessary concepts from functional data analysis. Section 3 explores Cramér-von Mises tests in a scalar setting and Section 4 introduces the theoretical foundations of permutation testing. Then, Section 5 focuses on the test developed in Bugni and Horowitz 2021 for the case of a two-sample test. The main contribution of this thesis is a variant of the test presented in Bugni and Horowitz 2021 with the goal to test for specific violations of the null hypothesis relating to the persistence of the data generating processes. This variant of the original test is presented in Section 6. Section 7 presents a set of simulations exploring the properties of the original test and shows a heuristic simulation to motivate the proposed extension. Section 8 applies the test from Bugni and Horowitz 2021 to half-hourly electricity demand data from Adelaide. Finally, Section 9 gives an outlook on possible further extensions, addresses some problems and shortcomings of the presented results, and outlines simulations that could be performed to better understand the properties of the procedures described in this thesis.

Two additional aspects should be mentioned. First, all code and data used in this thesis is available in the following public GitHub repository: `https://github.com/JakobJuergens/Masters_Thesis`. Second, Appendix 11.1 gives a very informal description of the test scenario and the underlying idea of the test. It shall serve as a primer for readers without prior knowledge of functional data analysis and permutation tests.

2 Functional Data Analysis

The overarching concept of functional data analysis is to analyze data that are functional in nature. In this context, functional observations can often be understood as smooth curves and functional data sets, thus, consist of realizations of processes that generate smooth curves. A classical example for a functional data set is shown in Figure 1. It depicts growth curves of 93 humans up to the age of 18 provided by the R package *fda*[1].

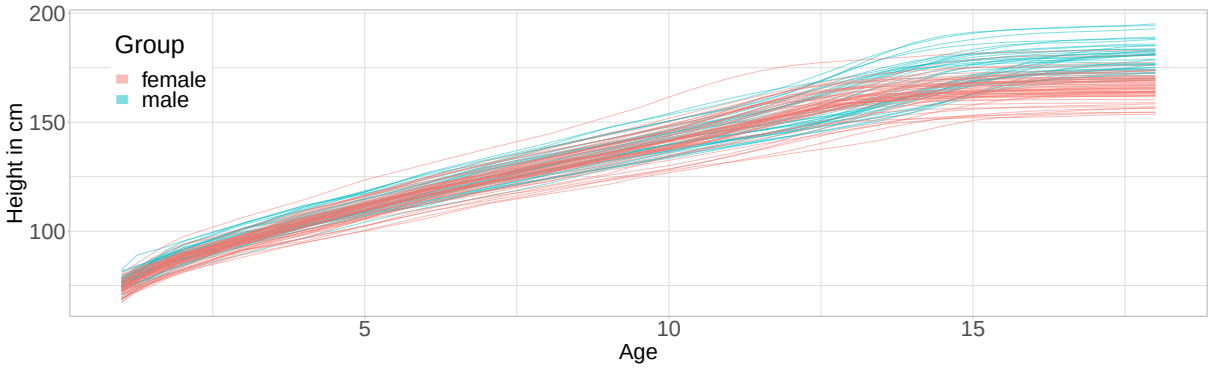

Figure 1: Human Growth Curves up to the Age of 18

Even though measurements are only taken at discrete ages, it is clear that each human has a height at every point in time. Thus, discrete data are only measurements of these underlying continuous curves. The higher the measurement frequency, the closer the data resemble the curves themselves. In many cases, functional data analysis restricts its scope to observations in specific subsets of the functions $f : \mathbb{R} \to \mathbb{R}$. As these are inherently infinite-dimensional, it is beneficial to introduce theory to make use of their unique properties. Sections 2.1 and 2.2 give an introduction to the necessary concepts closely following Hsing and Eubank 2015. More information on specific topics such as potential applications or the implementation in statistical software is given in Ramsay and Silverman 2005 and Kokoszka and Reimherr 2021.

2.1 Hilbert Space of Square Integrable Functions

For many methods, such as functional linear regression, it is beneficial to put additional restrictions on the analyzed functions. One typical assumption is that curves belong to the space of square-integrable functions on a compact interval, which is a so-called separable Hilbert space whose properties often simplify theoretical considerations in functional data analysis. As Bugni and Horowitz 2021 assume square-integrability, the following section introduces the concepts around Hilbert spaces in general and the space of square-integrable functions in particular.

To define Hilbert spaces, it is necessary to first define inner product spaces as Hilbert spaces are special cases of this larger category of spaces.

[1] Ramsay, Graves, and Hooker 2021.

Definition 2.1 (Inner Product)
A function $\langle \cdot, \cdot \rangle : \mathbb{V}^2 \to \mathbb{F}$ on a vector space $\mathbb{V}$ over $\mathbb{R}$ is called an inner product if the following four conditions hold for all $v, v_1, v_2 \in \mathbb{V}$ and $a_1, a_2 \in \mathbb{R}$.

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

Definition 2.2 (Inner Product Spaces, Orthogonality, and Hilbert Spaces)
A vector space with an associated inner product is called an inner product space. Two elements v_1 and v_2 of an inner product space are orthogonal if $\langle v_1, v_2 \rangle = 0$. An inner product space that is complete with respect to the distance induced by the norm $\|v\| = \sqrt{\langle v, v \rangle}$ is called a Hilbert space.

As previously mentioned, Hilbert spaces play an important role in functional data analysis, and many methods, such as functional linear regression, make extensive use of their properties. In the context of this thesis, they are especially useful to describe objects such as random functions. Analogously to vector spaces, it is useful to express elements of a Hilbert space as linear combinations of a basis. To do so, the classical idea of a finite basis used to express every element has to be extended. Hsing and Eubank 2015 define the closed span and orthonormal sequences in the following way, leading to a subsequent definition of bases for infinite-dimensional Hilbert spaces.

Definition 2.3 (Closed Span)
The closed span of a subset A of some normed space, e.g. a Hilbert space, is the closure of $span\,(A)$ with respect to the distance induced by the norm of the space. In the following it is denoted by $\overline{span\,(A)}$.

Definition 2.4 (Orthonormal Sequence in a Hilbert Space)
Let $\{x_n\}$ be a countable collection of elements in a Hilbert space such that every finite subcollection of $\{x_n\}$ is linearly independent. Define $e_1 = \frac{x_1}{\|x_1\|}$ and $e_i = \frac{v_i}{\|v_i\|}$ for

$$v_i = x_i - \sum_{j=1}^{i-1} \langle x_i, e_j \rangle e_j.$$

Then, $\{e_n\}$ is an orthonormal sequence and $\overline{span\,(\{x_n\})} = \overline{span\,(\{e_n\})}$

These definitions lead to a natural analogon of the basis in finite-dimensional vector spaces for the case of potentially infinite-dimensional Hilbert spaces.

Definition 2.5 (Orthonormal Basis of a Hilbert Space)
An orthonormal sequence $\{e_n\}$ in a Hilbert space $\mathbb{H}$ is called an orthonormal basis of $\mathbb{H}$ if $\overline{span\,(\{e_n\})} = \mathbb{H}$. A Hilbert space that possesses a countable complete orthonormal basis is called separable Hilbert space.

Separable Hilbert spaces form a special class of Hilbert space. Their properties often allow for significant simplifications in the derivation of theoretical properties of methods in functional data analysis. One important property is that a countable complete orthonormal basis $\{e_n\}$ can be used to express elements of the separable Hilbert space as shown in Equation 1 which will be used extensively in later parts of this thesis. One such case is the calculation of specific test statistics at the core of Bugni and Horowitz 2021.

$$\forall v \in \mathbb{H} \quad \exists \{a_k\}_{k=1}^{\infty} \quad \text{s.t.} \quad v = \sum_{k=1}^{\infty} a_k e_k \tag{1}$$

The most prevalent separable Hilbert space in functional data analysis is the space of square-integrable functions on a compact interval. However, it is often interesting to consider whether square-integrability is a necessary assumption. Due to its prevalence in functional data analysis, square-integrability is sometimes assumed even though generalizations to less restrictive function spaces might be possible.

Definition 2.6 (Hilbert Space of Square Integrable Functions)
The space of square-integrable functions on a compact interval $\mathcal{I} \subset \mathbb{R}$ together with the norm $\langle f, g \rangle = \int_{\mathcal{I}} f(t)g(t)\, \mathrm{d}t$ is a Hilbert space, denoted by $\mathbb{L}_2(\mathcal{I})$. A function $f : \mathcal{I} \to \mathbb{R}$ is called square-integrable if the following condition holds.

$$\int_{\mathcal{I}} [f(t)]^2 \, \mathrm{d}t < \infty \tag{2}$$

It is defined as a space of equivalence classes, where two functions are equivalent if they differ at most on a set of Lebesgue-measure zero.

Without loss of generality it is possible to reduce theoretical considerations to the case of $\mathcal{I} = [0, 1]$. Therefore, all treatment in this thesis focuses on the case of the unit interval. Atypically, Bugni and Horowitz 2021 define two square-integrable functions to be distinct if they differ on a non-empty set of Lebesgue-measure zero. However, as will become apparent in later parts of this thesis, using this definition in the context of Bugni and Horowitz 2021 only adds alternatives against which the test cannot have any power by construction. In most parts of this thesis, it is, therefore, sufficient to consider only the classical definition. However, in some places, it is necessary to use the definition from Bugni and Horowitz 2021, and the relevant differences will be explicitly named in those cases. An more general overview of the differences between the definitions is provided in Appendix 11.2.

2.2 Bases of $\mathbb{L}_2[0,1]$

One commonly used orthonormal basis of $\mathbb{L}_2[0,1]$ is the Fourier Basis. It consists of a series of functions $\left(\psi_i^F(x)\right)_{i=1}^{\infty}$ taken from the sine-cosine form of the famous Fourier series.

$$\psi_i^F(x) = \begin{cases} 1 & \text{if} \quad i = 1 \\ \sqrt{2}\cos(\pi i x) & \text{if} \quad i \quad \text{is even} \\ \sqrt{2}\sin(\pi(i-1)x) & \text{otherwise} \end{cases} \tag{3}$$

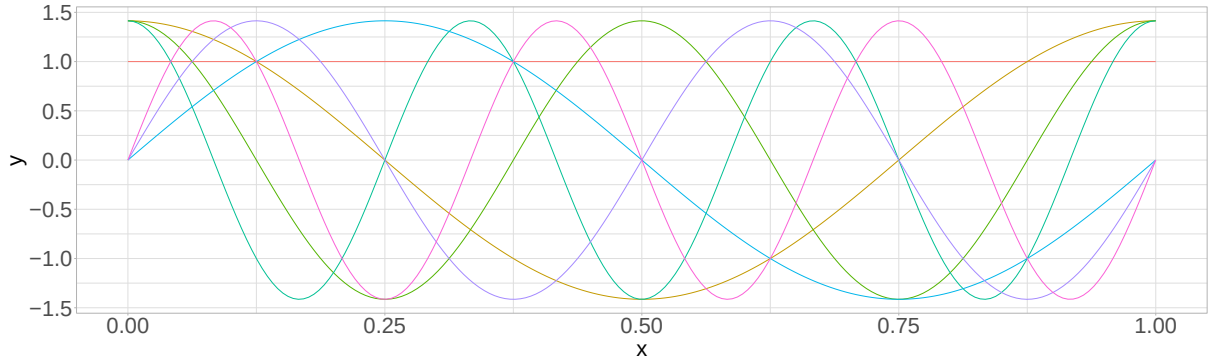

Figure 2: The first seven Fourier basis functions

A proof that the Fourier basis is a complete orthonormal basis of $\mathbb{L}_2[0,1]$ is given in section 2.4 of Hsing and Eubank 2015. As the Fourier basis is a countable orthonormal basis of $\mathbb{L}_2[0,1]$, this proof also shows that $\mathbb{L}_2[0,1]$ is a separable Hilbert space. Many other bases are often used in functional data analysis, such as b-splines and orthogonal polynomials. In most cases, the choice of the basis is motivated by the desired application, and the Fourier basis is typically chosen when the underlying process is assumed to generate periodical functions.

2.3 Random Functions

As a next step, it is important to consider the processes that generate functional observations. In statistics, it is common to interpret data as realizations of random variables, and in functional data analysis, the same holds true. However, since the observations are curves, the objects of interest are random variables that generate functional realizations. These are called random functions, and they are a special case of random variables. Paraphrasing from Bauer 2011 a general random variable is defined in the following way.

Definition 2.7 (Random Variable)
Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\Omega', \mathcal{A}')$ be a measure space. Then every $\mathcal{A}$-$\mathcal{A}'$-measurable function $X : \Omega \to \Omega'$ is called a $(\Omega', \mathcal{A}')$-random variable.

Since random functions are a special case, it is possible to define them by specializing the objects in Definition 2.7.

Definition 2.8 (Random Function)
A random variable that realizes in a function space with a corresponding σ-algebra, e.g. $\mathbb{L}_2[0,1]$ with its associated Borel σ-algebra, is called a random function.

2.4 Probability Measures on $\mathbb{L}_2[0,1]$

For the test developed in Bugni and Horowitz 2021, it is important to evaluate expectations of a functional on $\mathbb{L}_2[0,1]$ with respect to some probability measure μ. Thus, it is necessary to explore probability measures on $\mathbb{L}_2[0,1]$. However, due to the way measures are constructed in the paper, this description can be limited to probability measures induced by existing random functions. To understand how a random variable induces a measure, it suffices to look at a relatively basic fact from probability theory. Paraphrasing from Bauer 2011, let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}')$ be a random variable, then a probability measure $\mathbb{P}_X(B)$ on $\mathcal{A}'$ is induced by X as shown in Equation 4.

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)) \quad \forall B \in \mathcal{A}' \tag{4}$$

This idea specializes to the case of random functions realizing in $\mathbb{L}_2[0,1]$ and thereby introduces the option to use measures induced by existing random functions. Therefore, assume that $Z(t)$ is a random function in $\mathbb{L}_2[0,1]$ inducing a measure μ. As illustrated by Equation 1, $Z(t)$ can be expressed in terms of a functional basis. However, as $Z(t)$ is inherently random the corresponding Fourier coefficients $(b_k)_{k=1}^{\infty}$ are random as well.

$$Z(t) = \sum_{k=1}^{\infty} b_k \psi_k(t) \quad \text{s.t.} \quad \sum_{k=1}^{\infty} b_k^2 < \infty \tag{5}$$

The latter property is needed to ensure the square-integrability of the random function. As shown by this equation, a random function in $\mathbb{L}_2[0,1]$ is fully described by the distribution of a countably infinite, square-summable sequence of scalar random variables b_k. A general treatment of the theory of random sequences is out of the scope of this thesis. However, it is interesting to look at the relationship between the measure μ induced by the infinite sum and the measure μ_K induced by its finite approximation shown in Equation 6.

$$Z_K(t) = \sum_{k=1}^{K} b_k \psi_k(t) \tag{6}$$

Understanding the relationship between these measures justifies using a finite number of basis functions in the implementation of the test. Bugni, Hall, et al. 2009 show that the measure μ_K induced by $Z_K(t)$ converges to the measure μ induced by $Z(t)$ in the following sense: Let $\mathbf{A}$ be a μ-measurable subset of $\mathbb{L}_2[0,1]$. Then $\forall a \in \mathbf{A}$, there is a unique countably infinite sequence of Fourier coefficients $b(a) = (b_k(a))_{k=1}^{\infty}$ given by $b_k(a) = \int_0^1 a(t)\psi_k(t)\mathrm{d}t$. As $a \in \mathbb{L}_2[0,1]$ these coefficients fulfill $\sum_{k=1}^{\infty} b_k^2 < \infty$.

To define the mode of convergence Bugni, Hall, et al. 2009 define the following objects.

- $\mathbf{B} = \{b(a) \mid a \in \mathbf{A}\}$
- $\mathbf{B}_K = \{(b_1, \ldots, b_K) \mid \exists b' \in B \; \forall k = 1, \ldots, K \quad b_k = b'_k\}$
- $\mathbf{A}_K = \{\sum_{k=1}^{\infty} b_k \psi_k(t) \mid (b_1, \ldots, b_K) \in \mathbf{B}_K \quad \text{and} \quad \sum_{k=1}^{\infty} b_k^2 < \infty\}$

The authors then show that for any $\mathbf{A}$ in the Borel sigma algebra of subsets of $\mathbb{L}_2[0,1]$, the following condition holds, where $\mathbb{P}$ denotes the probability measure associated with μ.

$$\lim_{K \to \infty} \mathbb{P}(\mathbf{A}_K) = \mathbb{P}(\mathbf{A}) \tag{7}$$

This argument allows using a truncated basis in the construction of a measure at a reasonably high truncation parameter K as integrals with respect to μ_K over a functional $\mathcal{F}(z)$ converge to their counterparts with respect to μ as K goes to infinity.

$$\lim_{K \to \infty} \int_{\mathbb{L}_2[0,1]} \mathcal{F}(z) \mathrm{d}\mu_K(z) = \int_{\mathbb{L}_2[0,1]} \mathcal{F}(z) \mathrm{d}\mu(z) \tag{8}$$

A more detailed treatment of these aspects is outside the scope of this thesis. Therefore, this section shall primarily serve to give intuition. For readers interested in a more detailed mathematical analysis, Gihman and Skorokhod 2004 and Skorohod 1974 give a rigorous treatment of the necessary theory on measures, probabilities, and integration in Hilbert spaces.

2.5 Functional Integration on $\mathbb{L}_2[0,1]$

To evaluate the expected value mentioned at the beginning of Section 2.4, it is necessary to integrate over a function space. Therefore, exploring the ideas of functional integration and integration on separable Hilbert spaces is essential. The formal definition of a functional integral over the functional $G[f]$ over $\mathbb{L}_2[0,1]$ is given by Equation 9 as a nesting of countably many integrals over the joint distribution of the Fourier coefficients.

$$\int_{\mathbb{L}_2[0,1]} G\left[f\right] \left[Df\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G\left(f_1, f_2, \ldots\right) \prod_n \mathrm{d}f_n \tag{9}$$

An in-depth treatment of integration on Hilbert spaces is out of the scope of this thesis. A detailed overview of integration on Hilbert spaces is given in Skorohod 1974 and could be used to derive theoretical properties in a more general setting. Functional integrals over Hilbert spaces often do not have closed-form solutions. Instead, it is necessary to evaluate them using methods from perturbation theory. An introduction to the necessary theory is given in Jeribi 2021.

Because of the nature of the test in Bugni and Horowitz 2021, using these methods for approximation would be inadvisable due to the computational cost. Nevertheless, perturbation theory could be an interesting approach for theoretical considerations on the properties of Cramér-von Mises type tests in a functional setting. Instead, Bugni and Horowitz 2021 use Monte-Carlo integration to calculate a test statistic that is based on a functional integral. At this point, it is, thus, more beneficial to introduce Monte-Carlo integration with respect to some measure induced by a random variable. To illustrate the principle, assume that the following integral has a fixed value in $\mathbb{R}$ and that μ is induced by a continuous, scalar-valued random variable Z realizing in $[0, 1]$.

$$V = \int_0^1 g(x)\mathrm{d}\mu(x) = \int_0^1 g(x) f_Z(x)\mathrm{d}x \quad \text{where} \quad g : [0,1] \to \mathbb{R} \tag{10}$$

Assuming that it is possible to draw independent realizations of Z, an intuitive approach to approximate V is given by the Equation 11.

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \quad \text{where} \quad Z_i \sim_{\text{i.i.d.}} Z \quad i = 1, \ldots, n \tag{11}$$

This approach is justified because $\hat{V} \to_p V$ by the law of large numbers. Therefore, assuming that the expectation is finite, this approach can give approximations of the expectation that converge to the correct value as the number of realizations of Z goes to infinity. As the principle of this method does not depend on the domain of g, it can analogously be used in the context of function spaces. It gives a convenient alternative to the theoretically complex methods mentioned before and is, therefore, a useful alternative in the context of this thesis. More information on Monte-Carlo methods in general and, in particular, Monte-Carlo integration can be found in Shonkwiler and Mendivil 2009.

3 Cramér-von Mises Tests

The second prerequisite to understanding the test presented in Bugni and Horowitz 2021 is some knowledge about two-sample Cramér-von Mises (abbreviated as CvM in the following) type tests. To understand their working principle, it is useful to start in a simple setting, which is why this section is limited to the case of scalar random variables.

As mentioned in the introduction, it is often interesting to ask whether the same stochastic process generated the observations in two distinct data sets. In an experimental setting, one could ask whether a treatment assigned at random to a subset of agents changed the distribution of an outcome variable. The two-sample CvM test is one approach to answering this question.

3.1 Empirical Distribution Functions

The idea of CvM type tests is to compare the distribution functions of the underlying random variables. However, as these are typically unknown, it is necessary to define a sample analog to construct a suitable test statistic. Gibbons and Chakraborti 2021 define the order statistic of an observation in a sample and the empirical distribution function in the following ways, allowing the construction of a test statistic.

Definition 3.1 (Order Statistic)
Let $\{x_i \mid i = 1, \ldots, n\}$ be a random sample from a population with continuous cumulative distribution function F_X. Then there almost surely exists a unique ordered arrangement within the sample.

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$$

Then, $x_{(r)} \quad r \in \{1, \ldots, n\}$ is called the rth-order statistic.

Definition 3.2 (Empirical Distribution Function)
Assuming a continuously distributed random variable where there are no binds in a sample $\{x_i \mid i = 1, \ldots, n\}$ with probability one, the Empirical Distribution Function derived from the sample is defined as the following step function.

$$\hat{F}_n(x) = \begin{cases} 0 & \text{if} \quad x < x_{(1)} \\ \frac{r}{n} & \text{if} \quad x_{(r)} \leq x < x_{(r+1)} \\ 1 & \text{if} \quad x \geq x_{(n)} \end{cases} \tag{12}$$

An example showing the empirical distribution function for a selection of samples drawn from a standard normal distribution is given in Figure 11 in Appendix 11.8. As this plot illustrates, the empirical distribution function gets increasingly close to the theoretical distribution function of the underlying random variable as the number of realizations increases. In fact, the strong law of large implies that $\hat{F}_n(x) \xrightarrow{\text{a.s.}} F(x) \quad \forall x$ implying that $\hat{F}_n(x)$ is a consistent estimator of $F(x)$.

3.2 Null Hypothesis

As previously mentioned, the null hypothesis of CvM tests relates to distributional equality of random variables. To formulate the precise null hypothesis, let $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ be two data sets generated by continuous random variables $X \sim_{\text{i.i.d.}} F(t)$ and $Y \sim_{\text{i.i.d.}} G(t)$. The null hypothesis that both samples were generated by random variables following the same cumulative distribution function is then formulated as follows.

$$\begin{aligned} H_0 &: F(t) = G(t) \quad \forall t \in \mathbb{R} \\ H_1 &: \exists t \in \mathbb{R} \quad \text{s.t.} \quad F(t) \neq G(t) \end{aligned} \tag{13}$$

3.3 Two-Sample Cramér-von Mises Statistic

The CvM test is motivated by the idea that differences between the empirical distribution functions should be small under the null hypothesis. The test statistic is based on the integrated squared difference of the empirical distribution functions to quantify the difference between two functions. Adapting a definition from Büning and Trenkler 2013, it is defined as shown in Equation 14.

$$C_{m,n} = \frac{nm}{n+m} \int_{-\infty}^{\infty} \left(\hat{F}_m(t) - \hat{G}_n(t) \right)^2 \, \mathrm{d} \left(\frac{m\hat{F}_m(t) + n\hat{G}_n(t)}{m+n} \right) \tag{14}$$

Darling 1957 gives an overview on the CvM test and some related tests in both the one-sample goodness-of-fit and the two-sample setting. Anderson 1962 explores the small sample distribution of the two-sample test statistic and provides a comparison to the limiting distribution derived by Rosenblatt 1952 and Fisz 1960 which is given in Section 11.3 in the Appendix. These distributions can be used to perform tests with appropriate critical values and have been implemented in many packages for statistical computing.

It is possible to generalize this test by introducing a weight function $w(t)$ as shown in Equation 15.

$$T_{m,n}^{EDF} = \frac{nm}{n+m} \int_{-\infty}^{\infty} \left(\hat{F}_m(t) - \hat{G}_n(t) \right)^2 w(t) \, \mathrm{d} \left(\frac{m\hat{F}_m(t) + n\hat{G}_n(t)}{m+n} \right) \tag{15}$$

The CvM test corresponds to $w(t) = 1$. It is advisable to think about the desired properties of the test when choosing a specific weight function as it can affect the power against different types of alternatives. A well-known weight function for the one-sample CvM goodness of fit test was proposed by Anderson and Darling 1952 and Anderson and Darling 1954 and adapted to the two-sample setting by Pettitt 1976. The latter is given by $w(t) = \left[\left(\frac{m\hat{F}_m(t)+n\hat{G}_n(t)}{m+n} \right) \left(1 - \frac{m\hat{F}_m(t)+n\hat{G}_n(t)}{m+n} \right) \right]^{-1}$ and has been shown to have better properties in some scenarios as it assigns higher weight to observations that correspond to the tails of the distribution.

4 Permutation Tests

The third component to understanding Bugni and Horowitz 2021 is understanding the principles of permutation testing. In layman's terms, the idea of a permutation test is the following: if two samples show distinctly different properties, that will lead to differences in an appropriately chosen summary statistic. If one were to permute the elements of the samples randomly, one would expect these differences in the summary statistics to disappear. Permutation tests formalize this intuition. The following section closely follows chapter 15 from Lehmann and Romano 2005.

4.1 Functional Principle of Permutation Tests

One of the defining features of a test is its null hypothesis. In the case of randomization tests, which are a generalization of permutation tests, it is often formulated in the sense of distributional equality. This makes the approach particularly interesting for the questions studied in this thesis.

Let X be a random variable realizing in a sample space $\mathcal{X}$. Then, the null hypothesis in randomization testing is that the distribution P generating X belongs to a family of distributions Ω_0. For this very general setting, Lehmann and Romano 2005 define an assumption called the randomization hypothesis that allows for the construction of randomization tests.

Assumption 4.1 (Randomization Hypothesis)
Let G be a finite group of transformations $g : \mathcal{X} \to \mathcal{X}$. Under the null hypothesis of the randomization test, the distribution of X is invariant under the transformations $g \in G$. In other words, gX and X have the same distribution whenever X has distribution $P \in \Omega_0$.

For example, recall the scalar CvM test presented in Section 3. Under the null hypothesis, both X and Y follow the same distribution. As the random samples generated by both random variables are assumed to be independent, permuting the observations within and across samples will not change their joint distribution. Therefore, in this simple setting, the randomization hypothesis is fulfilled. The same argument trivially applies in the functional setting, justifying the use of a permutation test in Bugni and Horowitz 2021.

Under the randomization hypothesis one can construct a permutation test based on any test statistic $T : \mathcal{X} \to \mathbb{R}$ that is suitable to test the null hypothesis under consideration. Suppose that G has M elements, then given $X = x$, let

$$T_{(1)}(x) \leq T_{(2)}(x) \leq \cdots \leq T_{(M)}(x)$$

be the ordered values of the test statistic $T(gx)$ as described in Definition 3.1 as g varies over G. For a fixed nominal level $\alpha \in (0, 1)$, define $k = M - \lfloor M\alpha \rfloor$ and the following objects.

$$M^+ = \sum_{m=1}^{M} \mathbb{1}\left[T_{(m)}(x) > T_{(k)}(x)\right] \qquad M^0 = \sum_{m=1}^{M} \mathbb{1}\left[T_{(m)}(x) = T_{(k)}(x)\right] \tag{16}$$

Using these objects, it is possible to define a test function that decides whether the null hypothesis is rejected based on the value of the test statistic calculated on the non-permuted samples $\mathbf{T} = T(x)$ and a chosen size α for the test.

Definition 4.1 (Randomization Test Function)
The Randomization Test Function is a function $\phi : \mathcal{X} \to \mathbb{R}$ defined in the following way.

$$\phi(x) = \begin{cases} 1 & \text{if } \mathbf{T} > T_{(k)}(x) \\ a & \text{if } \mathbf{T} = T_{(k)}(x) \\ 0 & \text{if } \mathbf{T} < T_{(k)}(x) \end{cases} \quad \text{where} \quad a = \frac{M\alpha - M^+(x)}{M^0(x)}$$

This function can be interpreted as a decider of whether the corresponding null hypothesis is rejected. If $\phi(X) = 1$, the null hypothesis is rejected, if $\phi(X) = 0$ it cannot be rejected, and if $\phi(X) = a$, the decision is randomized using a Bernoulli variable that is 1 with probability a. Under Assumption 4.1, it is possible to show that, given a test statistic $T(X)$, the resulting test ϕ has size α.

$$\mathbb{E}_P\left[\phi(X)\right] = \alpha \quad \forall P \in \Omega_0 \tag{17}$$

This effectively means that a randomization test constructed in the described way always has the correct specified size. However, the power against alternatives can be highly sensitive to both the sample sizes and the number of transformations M. Similarly, the p-value of a randomization test given $X = x$ is calculated as shown in Equation 18.

$$\textit{p-value} = \frac{1}{M} \sum_{g \in G} \mathbb{1}\left[T(gx) \geq T(x)\right] \tag{18}$$

Lehmann and Romano 2005 explore the example of testing for the equality of the generating probability laws of two independent samples of scalar random variables. This is precisely the relevant application for the permutation variant of the two-sample CvM test that Bugni and Horowitz 2021 extend to the setting of functional data. In this two-sample setting, the transformation is applied to the combined sample. The transformed sample is then split again forming two samples to fit the structure of the original data set. Up to this point, the transformations $g \in G$ are still very general. In most scenarios, it is useful to specialize the kinds of transformations that are applied to the data to the permutations of the sample under consideration.

Definition 4.2 (Permutation)
Let S be a set, then a permutation of S is a bijective function $\pi : S \to S$.

If S is a finite set with N elements, there are $N!$ different permutations. Applying this idea to the setting of two samples with n and m observations respectively, there are $(n+m)!$ permutations in the combined set of observations. One way of describing the corresponding group of transformations G is shown in Equation 19.

$$\begin{aligned} \Pi_N &= \{\pi : \{1, \ldots, N\} \to \{1, \ldots, N\} \mid \pi \text{ is bijective}\} \\ G &= \left\{g : \mathbb{R}^N \to \mathbb{R}^N \mid \exists \pi \in \Pi_N \ \forall x \in \mathbb{R}^N \ g(x) = \left(x_{\pi(1)}, \ldots, x_{\pi(N)}\right)\right\} \end{aligned} \tag{19}$$

Using G as described in Equation 19 in the construction of a randomization test leads to the concept of permutation tests. These are special cases of the testing procedure introduced before and rely on the specific choice of G. Assuming that the test statistic is invariant over intra-sample permutations, it is possible to obtain identical results using combinations instead of permutations. As the number of combinations is $\binom{m+n}{m}$ and thereby considerably smaller than the number of permutations, this approach can be beneficial. In reality, however, it is often infeasible to calculate a test statistic on all M permuted samples. Therefore, it is typical to randomly sample a chosen number B of permutations used to approximate the corresponding values. These stochastic approximations are justified for large numbers of sampled permutations, as the corresponding p-values converge in probability to the p-value associated with the test using all permutations.
For the implementation of the test described in Bugni and Horowitz 2021 used in this thesis, distinct combinations are chosen if it is feasible to calculate the test statistic for all combinations. If this approach is infeasible due to the number of combinations, the implementation resorts to randomly sampling permutations with replacement instead.

4.2 Asymptotic Properties

One interesting question concerning the properties of permutation tests is their behavior when the number of observations goes to infinity. This section explores these properties, closely following 15.2.2 from Lehmann and Romano 2005. To study the asymptotic properties of permutation tests, it is useful to define the so-called randomization distribution.

Definition 4.3 (Randomization Distribution)
The randomization distribution of a test statistic T is defined by the following distribution function where $M = |G|$ is the number of transformations in G.

$$\hat{R}(t) = \frac{1}{M} \sum_{g \in G} \mathbb{1}\left[T(gX) \leq t\right]$$

It describes the distribution of the test statistic T under a uniformly distributed choice of transformation g. As X is itself a random variable, $\hat{R}(t)$ is random too.

It is additionally beneficial to consider not only a sequence of samples increasing in size but a sequence of settings where the prerequisite objects change accordingly. Consider therefore a sequence of settings with $X = X^N$, $P = P_N$, $\mathcal{X} = \mathcal{X}_N$, $G = G_N$, $T = T_N$ etc. for an increasing number of observations N. Defining the randomization distribution of T_N analogously to Definition 4.3 as $\hat{R}_N(t) = \frac{1}{M_N} \sum_{g \in G_N} \mathbb{1}\left[T_N(gX^N) \leq t\right]$ it is possible to analyze the limiting behavior of $\hat{R}_N(t)$ as $N \to \infty$ to understand the asymptotic properties of the permutation test. It is possible to show that for a random variable $\mathcal{G}_N$ that is uniformly distributed on G_N and independent of X^N, Equation 20 holds true under the randomization hypothesis.

$$\mathbb{E}[\hat{R}_N(t)] = \mathbb{P}\left[T_N(\mathcal{G}_N X^N) \leq t\right] = \mathbb{P}\left[T_N(X^N) \leq t\right] \tag{20}$$

Assuming that $T_N(X^N)$ converges in distribution to a stable limiting distribution with cumulative distribution function $R(t)$ which is continuous at t, it follows that $\mathbb{E}[\hat{R}_N(t)] \to R(t)$. Under some conditions it is additionally possible to show that $\hat{R}_N(t) \to_P R(t)$ at all continuity points t of $R(t)$. The proof of the latter statement presented in Lehmann and Romano 2005 courtesy of Hoeffding 1952 is given in Appendix 11.4.

It should be noted that to make use of the stable limiting distribution of, for example, the CvM test, as described in Appendix 11.3, it is necessary to let the ratio of observations in the two samples, n and m, go to a constant λ as the total number of observations N increases. A similar argument is necessary in the case of the CvM type test in Bugni and Horowitz 2021.

5 Test by Bugni and Horowitz (2021)

Similar to the idea of the CvM test in a scalar setting, Bugni and Horowitz 2021 devise a test for the equality of the distributions of the data generating processes of two independent random samples of functional data. To define the exact hypothesis, they define a distribution function for random variables realizing in $\mathbb{L}_2[0,1]$ as follows.

Definition 5.1 (Distribution Function of a Random Function)
Let $X : \Omega \to \mathbb{L}_2[0,1]$ be a random function realizing in the space of square-integrable functions. Then its distribution function is defined as the following object.

$$F_X(z) = \mathbb{P}\left[X(t) \leq z(t) \quad \forall t \in [0,1]\right] \quad z \in \mathbb{L}_2[0,1]$$

As previously mentioned, the authors assume that two functions $z_1, z_2 \in \mathbb{L}_2[0,1]$ are distinct even if they only differ on a set of Lebesgue-measure zero. In this case, this distinction is important to define the distribution function appropriately. This can be seen by checking whether the distribution function evaluated at two functions from the same equivalence class will necessarily evaluate to the same value, which is trivially not the case. Therefore, for the definition of the distribution function of a random variable in the sense of Bugni and Horowitz 2021, it is necessary to use their definition of $\mathbb{L}_2[0,1]$. However, as previously mentioned, this distinction will not have an influence in practice due to the construction of the test statistic and the chosen measure.

5.1 Assumptions

The authors explicitly make four assumptions, some of which can be relaxed to allow for a more general setting.

Assumption 5.1

1. $X(t)$ and $Y(t)$ are separable, μ-measurable stochastic processes
2. $\{X_i(t) \mid i = 1, \ldots, n\}$ is a random sample of the process $X(t)$.
   $\{Y_i(t) \mid i = 1, \ldots, m\}$ is a random sample of $Y(t)$ independent of $\{X_i(t) \mid i = 1, \ldots, n\}$.

As this assumption contains multiple points, it is useful to address them individually.

- μ is a measure chosen by the researcher, and its construction is explored in detail in Section 5.5.
- The assumptions from the scalar CvM test, independence of observations within sample, and independence between the samples, carry over to allow for the construction of a CvM type test in a functional setting.
- The CvM type test hinges on the measurability of the processes with respect to the chosen measure μ. Therefore, measurability with respect to μ must be fulfilled for the test to be applicable. As μ is a measure on $\mathbb{L}_2[0,1]$ this assumption entails that the processes have to realize in $\mathbb{L}_2[0,1]$.
- Separability of the stochastic processes is a less common assumption. Paraphrasing from Gihman and Skorokhod 2004, a separable stochastic process is defined as shown in Definition 5.2.

Definition 5.2 (Separable Stochastic Process)
A real-valued random function $X(t, \omega)$ – here denoted explicitly as a function of ω – with an associated probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is separable if:

1. There is a dense countable subset $I \subset T$ of its index set T
2. There is a set $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 0$ such that for an arbitrary open set $G \subset T$ and an arbitrary closed set $F \subset \mathbb{R}$ the two sets $\{\omega \mid X(t, \omega) \in F \quad \forall t \in G\}$ and $\{\omega \mid X(t, \omega) \in F \quad \forall t \in G \cap I\}$ differ from each other at most on Ω_0

In less theoretical terms, this means that the properties of the process are determined by its behavior on a countable subset of points of its index set. This assumption is helpful in proving some results on the asymptotic properties of the CvM statistic that rely on specific measurability properties. However, this assumption could potentially be relaxed if weaker restrictions on the processes are sufficient to provide the necessary properties.

Assumption 5.2
$\mathbb{E}[X(t)]$ and $\mathbb{E}[Y(t)]$ exist and are finite for all $t \in [0,1]$.

Assumption 5.2 is necessary for one of the two tests used in Bugni and Horowitz 2021. The test is based on the squared difference between the sample mean functions to detect differences in the respective mean functions. Therefore it relies on the existence and finiteness of the mean functions in every point of the domain.

Assumption 5.3
$X_i(t)$ and $Y_i(t)$ are observed for all $t \in [0, 1]$.

Assumption 5.3 can be relaxed, and a similar test can be constructed for the case of discretely observed processes. This variation of the test will not be addressed explicitly in this thesis. However, Bugni and Horowitz 2021 describe how to extend their idea to this common scenario. Some of the simulations presented in later parts of this thesis make use of this extension.

5.2 Null Hypothesis and Related Tests

As with most permutation tests, the null hypothesis of the test presented by Bugni and Horowitz 2021 is distributional in nature. In the following let $F(z)$ denote the distribution function of the random function $X(t)$ and $G(z)$ denote the distribution function of the random function $Y(t)$. Then the null hypothesis can be formulated in the following way.

$$\begin{aligned} H_0 &: \quad F(z) = G(z) \quad \forall z \in \mathbb{L}_2[0,1] \\ H_1 &: \quad \mathbb{P}_\mu \left[F(Z) \neq G(Z) \right] > 0 \end{aligned} \tag{21}$$

Here, μ is a probability measure on $\mathbb{L}_2[0, 1]$ and Z is a random function realizing in $\mathbb{L}_2[0, 1]$ with a distribution induced by μ. As can be seen, by the formulation of the hypothesis, this test does not intend to have power against alternatives that differ only on a set of functions with μ-measure zero. The choice of the measure is an important factor to consider when trying to use this test against a specific suspected alternative. Additionally, due to the way the measure is constructed and approximated, which is described in Section 5.5, alternatives that are based on functions that are identical almost everywhere always have μ-measure zero.

Tests for this hypothesis exist in some forms already, and the following list shall give a non-exhaustive overview of other procedures that could be applied. Hall and Tajvidi 2002 formulate a permutation test for high-dimensional data that can also be functional. Their permutation test is based on pairwise distances between observations and the number of observations from each sample that are smaller than a reference value. Hall and Van Keilegom 2007 develop a very similar approach to Bugni and Horowitz 2021 by using the integrated squared difference between estimates of the empirical distribution functions. However, their approach is not based on a permutation test but on bootstrapping to derive critical values for a test and specific smoothing approaches for the observed data. Additionally, their method uses the Karhunen-Loève expansion to construct a measure in a similar sense as presented in this thesis. Pomann, Staicu, and Ghosh 2016 develop a framework for discretely observed functional curves using a marginal functional principal components approach to reduce the problem to a finite number of dimensions. Their approach then makes use of finite-dimensional variants of tests such as the Anderson-

Darling test mentioned in Section 3. Corain et al. 2014 propose a test using so-called derived variables that are calculated from the data, such as, for example, the functional principal components. They reformulate the null hypothesis in terms of these derived variables and suggest a framework combining multiple dependent randomization tests for the individual hypotheses relating to the different derived variables.

5.3 Cramér-von Mises type Test

Similar to the CvM test described in Section 3, the test constructed by Bugni and Horowitz 2021 relies on the empirical distribution functions describing the samples. Therefore, it is necessary to define empirical distribution functions for the functional setting:

$$\hat{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big[X_i(t) \leq z(t)\ \forall t \in [0,1]\big] \qquad \hat{G}_m(z) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\big[Y_i(t) \leq z(t)\ \forall t \in [0,1]\big] \tag{22}$$

As in the scalar setting described in Section 3, the idea of the test is that the differences between these empirical distribution functions are expected to be comparatively small under the null hypothesis. This difference can be quantified using the integrated squared difference leading to the following equation.

$$\tau = \int_{\mathbb{L}_2[0,1]} [F(z) - G(z)]^2 \, \mathrm{d}\mu(z) \tag{23}$$

As the distribution functions are typically unknown, it is possible to construct a test statistic using the empirical distribution functions as shown in Equation 24.

$$\tau_{n,m} = (n+m) \int_{\mathbb{L}_2[0,1]} \left[\hat{F}_n(z) - \hat{G}_m(z)\right]^2 \mathrm{d}\mu(z) \tag{24}$$

As algebraic functional integration is unsuitable for this testing procedure due to the structure of the underlying objects, the authors use Monte-Carlo integration as explained in Section 2.5 for approximating this integral. To perform the test, it is, therefore, necessary to choose a parameter L that determines the number of functions $\{Z_l \mid l = 1, \ldots, L\}$ that is used to approximate the value of $\tau_{n,m}$ by the following equation.

$$\hat{\tau}_{n,m} = \frac{n+m}{L} \sum_{l=1}^{L} \left[\hat{F}_n(Z_l) - \hat{G}_m(Z_l)\right]^2 \tag{25}$$

These functions are drawn from a random function associated with the chosen measure μ, and their construction is explained in more detail in Section 5.5. At this point, it is interesting to ask whether Monte-Carlo integration is a suitable approach in this specific setting; in other words, whether the mean of $\left[\hat{F}_n(z) - \hat{G}_m(z)\right]^2$ with respect to μ exists and is finite.

Assuming existence, which depends on the chosen measure, an argument for finiteness is given in Appendix 11.5. Using the theory from Section 4 the critical values for the permutation test statistic are then calculated in the following way.

$$t^{*,\tau}_{n,m}(1-\alpha) = \inf\left\{ t \in \mathbb{R} \quad | \quad \frac{1}{Q}\sum_{q=1}^{Q} \mathbb{1}\left[\tau_{n,m,q} \leq t\right] \geq 1-\alpha \right\} \tag{26}$$

Here $\tau_{n,m,q}$ denotes the test statistic calculated on the samples permuted using permutation q, assuming that Q permutations were considered overall. With these critical values, it is now possible to perform the test at a chosen significance level α.

5.4 Asymptotics for the Cramér-von Mises type Test

Similar to the case presented in Bugni, Hall, et al. 2009, it is possible to derive an asymptotic distribution for the CvM type test. Even though this is not necessary to perform the described permutation test, it could be an interesting benchmark to compare the test procedure in future large sample simulations. The authors mention that under the null hypothesis and mild regularity conditions, $\frac{1}{\sqrt{n+m}}\left[\hat{F}_n(z) - \hat{G}_m(z)\right]$ converges to a mean zero Gaussian process $\Upsilon(z) \quad z \in \mathbb{L}_2[0,1]$. Here, $F(z)$ and $G(z)$ again denote the distribution functions corresponding to the samples and $\hat{F}_n(z)$ and $\hat{G}_m(z)$ their empirical equivalents. $\Upsilon(z)$ has the covariance function shown in Equation 27.

$$\begin{aligned} cov[\Upsilon(z), \Upsilon(\tilde{z})] &= \frac{(1+\lambda)^2}{\lambda}\left\{G(\min(z,\tilde{z})) - G(z)G(\tilde{z})\right\} \\ \text{where} \quad \min(z,\tilde{z})(t) &= \min\left(z(t), \tilde{z}(t)\right) \quad \forall t \in [0,1] \end{aligned} \tag{27}$$

As in the case of the limiting distribution of the scalar CvM test described in Appendix 11.3, λ denotes the limiting ratio of observations in the two samples as $N = n + m \to \infty$, meaning $\frac{n}{m} \to \lambda$. Thus, the following limiting distribution can be obtained for $\tau_{n,m}$.

$$\tau_{n,m} \to_d \int_{\mathbb{L}_2[0,1]} \Upsilon^2(z)\, \mathrm{d}\mu(z) \tag{28}$$

The derivation of this limiting distribution follows a similar argument as the proof of Theorem 3.1 given in Bugni, Hall, et al. 2009. It is presented in the supplementary material of Bugni and Horowitz 2021. Using the arguments presented in the proof and in Section 4.2, it is possible to show that this limiting distribution also applies to the permutation variant of the CvM type test described in this thesis. In the context of the limiting distribution, it is again necessary to consider the implications of the definition of $\mathbb{L}_2[0,1]$ as given by Bugni and Horowitz 2021. Even though this distinction is not considered in detail in this thesis, for theoretical considerations of this test, it is vital that, for example, $\Upsilon(z)$ is defined with respect to the definition of $\mathbb{L}_2[0,1]$ from Bugni and Horowitz 2021.

5.5 Construction of the Measure μ

As can be seen in the fact that the alternative is formulated with respect to a measure μ on $\mathbb{L}_2[0,1]$, this measure significantly impacts the test's performance. Compared to the scalar CvM test described in Section 3, it fulfills a similar purpose to the weight function described in Section 3.3 in that it determines against which alternatives the test has comparatively high power. It is, therefore, crucial to choose the measure according to the kind of violation of the null hypothesis that is suspected. If no specific violation is expected, a choice similar to a constant weight function can be made. In a scalar CvM type test, a constant weight function gives the same weight to the whole domain. Intuitively this corresponds to some uniform distribution on $\mathbb{L}_2[0,1]$ which does not exist. Instead, the original authors refer to a constant mean function of the inducing random variable. As hinted at in Section 2.4, Bugni and Horowitz 2021 approach the problem of choosing a measure by constructing a random function $Z(t)$ that induces a suitable probability measure μ. This random function is constructed as shown in Equation 29.

$$Z(t) = \sum_{k=1}^{\infty} b_k \psi_k(t) \quad \text{s.t.} \quad \sum_{k=1}^{\infty} b_k^2 < \infty \tag{29}$$

By truncating the random function to use a finite amount of basis functions, as shown in Equation 30, it is possible to draw realizations from it. As explained in Section 2.4, the measure induced by this truncated random variable converges to the measure induced by the full random variable as the truncation parameter K goes to infinity. This fact justifies the approach theoretically.

$$Z_K(t) = \sum_{k=1}^{K} b_k \psi_k(t) \tag{30}$$

The construction hinges on specifying appropriate distributions for the Fourier coefficients $(b_k)_{k=1}^K$. To obtain a reasonable degree of power, it is crucial to specify the distribution of the random function $Z(t)$ such that its realizations are able to distinguish potential differences in the samples. This could for example be done by choosing a function $w(t) : [0,1] \to \mathbb{R}$ specifying the expectation of $Z(t)$ dependent on the data. Calculating the expected values of the Fourier coefficients as shown in Equation 31 then leads to the expected value of the random function $Z_K(t)$.

$$\mathbb{E}[b_k] = \int_0^1 w(t)\psi_k(t)\, \mathrm{d}t \quad \text{leading to} \quad \mathbb{E}[Z_K(t)] = \sum_{k=1}^{K} \mathbb{E}[b_k]\, \psi_k(t) \tag{31}$$

Lastly, a distribution around the expected value has to be specified. In Equation 32 this takes the form of random errors U_k and non-stochastic square-summable factors $(\rho_k)_{k=1}^{\infty}$.

$$b_k = \mathbb{E}[b_k] + \rho_k U_k \quad \text{s.t.} \quad \sum_{k=1}^{\infty} \rho_k^2 < \infty \tag{32}$$

The authors suggest choosing $w(t)$ to be comparatively large where differences between processes in the two samples are expected to be large. However, if no prior exists, a choice has to be made nevertheless. The following mean functions are shown in Figure 3.

1. In their application, Bugni and Horowitz 2021 set the distribution of the Fourier coefficients manually, as $b_1 \sim \mathcal{N}(\mu_1, \frac{1}{K})$ and $b_k \sim \mathcal{N}(0, \frac{1}{K}) \quad k = 2, \ldots, K$ where

$$\mu_1 = median_i \left\{ \max_{t \in [0,1]} X_i(t) \mid i = 1, \ldots, n, \right\}$$

   This method is equivalent to setting $w(t) = \mu_1 \quad \forall t \in [0, 1]$, as the inner product of any constant function with Fourier basis functions of order two or higher will be zero due to their cyclical nature.

2. Another idea is to choose a non-constant weight function based on sample 1. One option is to use specific point-wise quantiles $w(t) = quantile_q\{X_i(t) \mid i = 1, \ldots n\}$. This allows the generated functions for the approximation of the CvM statistic to resemble the reference data more closely, which could improve the test's properties.

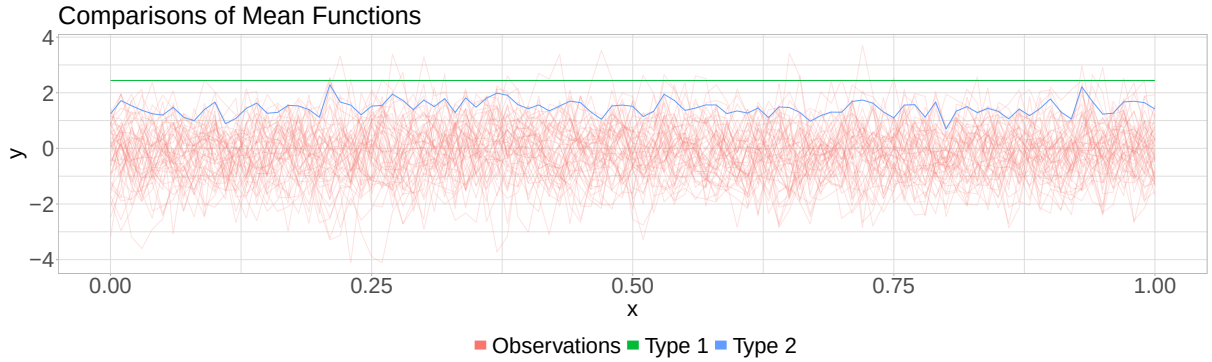

Figure 3: Mean Functions calculated for a Reference Sample

In addition to the mean function $w(t)$, the choice of the sequence $(\rho_k)_{k \in \mathbb{N}}$ and the distribution of the error terms U_k is of importance. The authors do not recommend a procedure for choosing $(\rho_k)_{k=1}^{\infty}$. However, it seems reasonable to choose parameters that lie in the vicinity of the standard deviations for the Fourier coefficients of the reference sample when fitting a Fourier basis to the original observations. Other methods, such as specifying a sequence without considering the data, are possible. For example, in the application part of Bugni and Horowitz 2021, the authors use $\frac{1}{K}$ for every Fourier coefficient. Concerning the choice of distribution for U_k, the authors recommend different distributions depending on the properties of the samples. The authors recommend one of the following choices for samples that indicate a thin-tailed distribution of $X(t)$.

- $U_k \sim \mathcal{N}(0, 1)$
- $U_k \sim Unif\left[-3^{\frac{1}{3}}, 3^{\frac{1}{3}}\right]$

If instead, the sample hints at a heavy-tailed distribution, one should consider a heavy-tailed distribution for the Fourier coefficients, such as a t-distribution with a low number of degrees of freedom.

5.6 Mean-focused Test

In addition to the CvM type test, the authors introduce a second test statistic aimed at detecting violations due to a mean shift in the data generating process. This test is added due to low power against pure mean shift alternatives of the CvM type test in the original paper's simulations. It is supposed to complement this weakness of the other test statistic. The variable ν on which this second test is based is the distance between the mean functions of the data generating processes as induced by the natural norm on $\mathbb{L}_2[0,1]$.

$$\nu = \int_0^1 \left(\mathbb{E}\left[X(t)\right] - \mathbb{E}\left[Y(t)\right]\right)^2 \mathrm{d}t \tag{33}$$

Replacing the expected values with their sample equivalents leads to a test statistic that can be used to perform a permutation test. The sample equivalents are defined as shown in Equation 34 and the resulting test statistic is shown in Equation 35.

$$\hat{\mathbb{E}}\left[X(t)\right] = \frac{1}{n}\sum_{i=1}^{n} X_i(t) \qquad \hat{\mathbb{E}}\left[Y(t)\right] = \frac{1}{m}\sum_{i=1}^{m} Y_i(t) \tag{34}$$

$$\nu_{n,m} = (n+m)\int_0^1 \left[\hat{\mathbb{E}}\left[X(t)\right] - \hat{\mathbb{E}}\left[Y(t)\right]\right]^2 \mathrm{d}t \tag{35}$$

In the same way the critical values for the CvM test are derived using the techniques described in Section 4.2, the critical values for the mean-based test are calculated as shown in Equation 36. Analogous to the previous case, $\nu_{n,m,q}$ denotes the test statistic calculated on the permuted samples corresponding to permutation q.

$$t^{*,\nu}_{n,m}(1-\alpha) = \inf\left\{ t \in \mathbb{R} \quad | \quad \frac{1}{Q}\sum_{q=1}^{Q} \mathbb{1}\left[\nu_{n,m,q} \leq t\right] \geq 1-\alpha \right\} \tag{36}$$

Similarly to the CvM type test, it is possible to show that $\nu_{n,m}$ weakly converges to a mean-zero Gaussian process as $n \to \infty$, $m \to \infty$ and $\frac{n}{m} \to \lambda$. As explained before, this limiting distribution carries over to the permutation test under the conditions outlined in Section 4.

At this point, it is interesting to mention that this test could be used on its own if the null hypothesis were limited to the equality of the mean functions of the data-generating processes. This could be done both using the permutation variant described in this thesis but also using the limiting distribution as described by the original authors. This is an interesting question in many applied contexts, and it would be interesting to study the individual test's properties compared to existing tests.

5.7 Combined Permutation Test

Using the test statistics described in Sections 5.3 and 5.6, it is possible to combine both into one test, trying to harness the advantageous properties of both tests. Therefore, define for the two underlying tests the following permutation test functions as described in Definition 4.1.

$$\phi_{n,m} = \begin{cases} 1 & \text{if } \tau_{n,m} > t^*_{n,m}(1-\alpha_\tau) \\ a_\tau & \text{if } \tau_{n,m} = t^*_{n,m}(1-\alpha_\tau) \\ 0 & \text{if } \tau_{n,m} < t^*_{n,m}(1-\alpha_\tau) \end{cases} \qquad \tilde{\phi}_{n,m} = \begin{cases} 1 & \text{if } \nu_{n,m} > t^*_{n,m}(1-\alpha_\nu) \\ a_\nu & \text{if } \nu_{n,m} = t^*_{n,m}(1-\alpha_\nu) \\ 0 & \text{if } \nu_{n,m} < t^*_{n,m}(1-\alpha_\nu) \end{cases} \tag{37}$$

Here, a_τ and a_ν are once again given by the following equations to ensure that the expected values of ϕ and $\tilde{\phi}$ have the desired values.

- $a_\tau = \frac{Q\alpha_\tau - Q^+_\tau}{Q^0_\tau}$
- $Q^+_\tau = \sum_{q=1}^{Q} \mathbb{1}\left[\tau_{n,m,q} > t^*_{n,m}(1-\alpha_\tau)\right]$
- $Q^0_\tau = \sum_{q=1}^{Q} \mathbb{1}\left[\tau_{n,m,q} = t^*_{n,m}(1-\alpha_\tau)\right]$
- $a_\nu = \frac{Q\alpha_\nu - Q^+_\nu}{Q^0_\nu}$
- $Q^+_\nu = \sum_{q=1}^{Q} \mathbb{1}\left[\nu_{n,m,q} > t^*_{n,m}(1-\alpha_\nu)\right]$
- $Q^0_\nu = \sum_{q=1}^{Q} \mathbb{1}\left[\nu_{n,m,q} = t^*_{n,m}(1-\alpha_\nu)\right]$

By the theoretical considerations of Section 4.2, it is clear that under Assumption 5.1 and 5.2, under the null hypothesis and for any $\alpha_\tau, \alpha_\nu \in (0,1)$, the following statements hold. This means that regardless of most choices that had to be made during the construction of the test, the size of the individual tests will be equal to the corresponding α.

$$\mathbb{E}_P\left(\phi_{n,m}\right) = \alpha_\tau \qquad\qquad \mathbb{E}_P\left(\tilde{\phi}_{n,m}\right) = \alpha_\nu \tag{38}$$

To combine the two tests, the authors use the well-known Bonferroni correction. It is used to adjust the individual tests' levels to control the family-wise error rate of the combined procedure, where the null hypothesis is rejected if either test indicates a violation at the adjusted level. Under H_0 this approach leads to the following relationship motivating a choice of $\alpha_\tau + \alpha_\nu = \alpha$ for a chosen overall size for the combined test of α. More information on the Bonferroni correction is given in Appendix 11.6.

$$\max(\alpha_\tau, \alpha_\nu) \leq \mathbb{P}\left[(\phi_{n,m} > 0) \cup (\tilde{\phi}_{n,m} > 0)\right] \leq \alpha_\tau + \alpha_\nu \tag{39}$$

One critique of this approach is that combined tests constructed using a Bonferroni correction are typically conservative. However, even though more advanced techniques can be used in some settings to combat this problem, this seems infeasible for the case of the test statistics presented in Bugni and Horowitz 2021. Additionally, this procedure introduces another choice as the values of α_τ and α_ν can be specified freely under the constraint that they add up to a chosen significance level for the combined procedure. In an applied setting, it is not trivial what the optimal choice of these values is.

5.8 Asymptotic Properties under the Alternative

As alluded to in Sections 3 and 5.3, for the asymptotic properties of the test, it is useful to make another assumption on the sequence of tests that is considered as the number of observations goes to infinity.

Assumption 5.4
As $N = m + n \to \infty$ let $\frac{n}{m} \to \lambda$ for some $\lambda \in \mathbb{R}_{>0}$.

This assumption allows the derivation of a stable limiting distribution and helps with the asymptotic properties of the test if the null hypothesis is incorrect. Under Assumptions 5.1 and 5.4, Bugni and Horowitz 2021 show that the following statement holds.

$$\int_{\mathbb{L}_2[0,1]} [F(z) - G(z)]^2 \, \mathrm{d}\mu(z) > 0 \implies \lim_{N\to\infty} \mathbb{P}\left[\tau_{n,m} > t^{*,\tau}_{n,m}(1-\alpha_\tau)\right] = 1 \quad \forall \alpha_\tau \in (0,1) \tag{40}$$

If additionally Assumption 5.2 holds, then the analogous statement holds for the mean-focused test.

$$\int_0^1 \left[\mathbb{E}\left[X(t)\right] - \mathbb{E}\left[Y(t)\right]\right]^2 \mathrm{d}t > 0 \implies \lim_{N\to\infty} \mathbb{P}\left[\nu_{n,m} > t^{*,\nu}_{n,m}(1-\alpha_\nu)\right] = 1 \quad \forall \alpha_\nu \in (0,1) \tag{41}$$

The combination of these statements can be used to show that the combined test is consistent.

5.9 Problematic Curve Fitting

One crucial question in implementing the method is whether to employ curve-fitting using a functional basis. It seems natural to fit the same basis to the data that is used for constructing the measure. Then the comparison between functions necessary for approximating the CvM type statistic is reduced to a problem of finding the zeroes of a finite Fourier-series, which could be approached with methods such as those presented in Boyd 2006. However, this can lead to problems illustrated by Figure 4, which shows curves that were fitted to samples generated by two data generating processes.

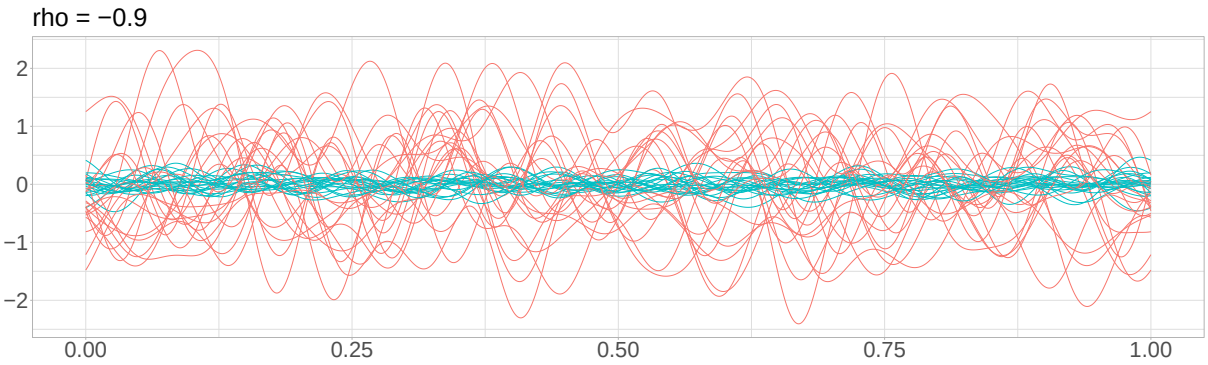

Figure 4: Artifact of the Curve Fitting Procedure
Sample 1 in red, Sample 2 in blue

These processes are constructed as described in Section 7 for a choice of $\rho_2(t) = -0.9$, ensuring that the data generating processes both exhibit a point-wise variance of one. However, the curve fitting procedure fails to conserve this property as high-frequency changes in the generated sample could be approximately orthogonal to the low-frequency basis functions contained in a truncated basis. Therefore, curve-fitting artifacts, such as those shown in Figure 4, can heavily influence the power of the constructed test, meaning that it could be advisable to change the approach to one of the following.

1. **Comparisons at the discrete data points.** Suppose observations are discrete but dense enough to be interpreted as smooth curves. In that case, one could use the points where data is observed to compare the functions generated for approximating the CvM test to the original samples. This approach is close to the one described in Bugni and Horowitz 2021 for the case of discrete observations.

2. **Linear Interpolation.** Instead, one could use linear interpolation of the original curves to compare them to the generated curves. This would preserve the functional nature while not generating the artifacts shown in Figure 4.

The simulations presented in Section 7 use the first approach as it limits the computational cost of the procedure while staying close to the original paper. The same problem occurs for the mean-focused test. In this case, linear interpolation between the discrete points is used to approximate the mean function on the full interval.

6 Test for Persistence Alternatives

One extension of the CvM type test presented in Bugni and Horowitz 2021 is a slight modification leading to a test for differences in the persistence structure of the data generating processes. This test is an interesting extension as few existing methods test a similar null hypothesis.

To give some context: In the literature, there is a considerable number of methods to test whether two samples share a common mean function. Cox and Lee 2008 propose a method that compares functions point-wise using a multiple comparison procedure introduced by Westfall and Young 1993. They show that this approach can be used to approximate a continuum of comparisons as the grid used to compare points gets finer. Fan and Lin 1998 develop a method based on the adaptive Neyman methodology to compare multiple samples. Their procedure uses wavelet thresholding techniques developed in the authors' earlier works. Lee, Cox, and Follen 2015 introduce a method that is also based on the adaptive Neyman methodology and a functional principal component approach to test whether two samples share a common mean function.
There are also tests for differences in the covariance functions of processes. Guo, Zhou, and Zhang 2018 develop a method based on the supremum of the integrated squared difference

between the sample covariance functions of the individual samples and the pooled sample. Test decisions are then made using critical values derived via bootstrapping approaches. Building on their previous work, Guo, Zhou, and Zhang 2019 propose an additional test that generalizes the idea of the point-wise F-test as described, for example, in Ramsay and Silverman 2005. In their approach, the authors extend existing ideas such as the F_{max}-test or the quasi-GPF trying to answer the corresponding global hypothesis to compare the covariance functions of multiple functional populations. However, few tests specifically address differences in the persistence of processes, including higher cross-moment functions and the proposed procedure adds to the limited arsenal.

6.1 Null-Hypothesis

Let $\{X_i(t) \mid i = 1, \dots, n\}$ and $\{Y_i(t) \mid i = 1, \dots, m\}$ again denote the samples under consideration. Assume that the samples consist of i.i.d. realizations of random functions $X(t)$ and $Y(t)$. To focus on the persistence properties of the random functions, define the standardized counterparts to these random functions in the following way.

$$\begin{aligned} \mu_X(t) =& \mathbb{E}\left[X(t)\right] & \mu_Y(t) =& \mathbb{E}\left[Y(t)\right] \\ \sigma_X(t) =& \sqrt{\mathbb{E}\left[\left(X(t) - \mu_X(t)\right)^2\right]} & \sigma_Y(t) =& \sqrt{\mathbb{E}\left[\left(Y(t) - \mu_Y(t)\right)^2\right]} \end{aligned} \tag{42}$$

Then the standardized random variables can be defined as follows.

$$\tilde{X}(t) = \frac{X(t) - \mu_X(t)}{\sigma_X(t)} \qquad \tilde{Y}(t) = \frac{Y(t) - \mu_Y(t)}{\sigma_Y(t)} \tag{43}$$

Their distribution functions can be defined analogously to Section 5 and shall be denoted by $F_{\tilde{X}}(z)$ and $F_{\tilde{Y}}(z)$ respectively in the following. Then a modified null hypothesis similar to Bugni and Horowitz 2021 can be formulated with respect to the distribution functions of the standardized random variables.

$$\begin{aligned} H_0 :& \quad F_{\tilde{X}}(z) = F_{\tilde{Y}}(z) \quad \forall z \in \mathbb{L}_2[0.1] \\ H_1 :& \quad \mathbb{P}_\mu\left[F_{\tilde{X}}(z) \neq F_{\tilde{Y}}(z)\right] > 0 \end{aligned} \tag{44}$$

As the standardization eliminates all differences in these processes' mean and variance patterns, one of the distinct remaining features that might differentiate these processes is their persistence properties. Somewhat more technical, this hypothesis is linked to the equality of the cross-moment functions of the underlying processes after standardization. Therefore, this focused approach could be tailored by choosing an appropriate μ to detect differences in this general idea of persistence of the data generating processes.

6.2 Test-Procedure

The modified test is functionally nearly identical to the CvM type test described in Bugni and Horowitz 2021. However, it adds a standardization step to eliminate differences in the mean functions of the processes and their variance structure. In each permutation step, the calculation of the test statistic is thus described by the following algorithm.

Let $\mathcal{W} = \{W_i(t) \mid i = 1, \dots, n\}$ be sample one and $\mathcal{V} = \{V_i(t) \mid i = 1, \dots, m\}$ be sample two. These samples are generated by randomly permuting the original samples $\{X_i(t) \mid i = 1, \dots, n\}$ and $\{Y_i(t) \mid i = 1, \dots, m\}$ across and within sample.

1. Calculate the sample mean functions for the permuted samples:
   $\bar{W}(t) = \frac{1}{n} \sum_{i=1}^{n} W_i(t)$ and $\bar{V}(t) = \frac{1}{m} \sum_{i=1}^{m} V_i(t)$
2. Center the sample curves to obtain centered permutation samples:
   $\bar{\mathcal{W}} = \{W_i(t) - \bar{W}(t) \mid i = 1, \dots, n\}$ and $\bar{\mathcal{V}} = \{V_i(t) - \bar{V}(t) \mid i = 1, \dots, m\}$
3. Calculate the point-wise standard deviations:
   $\sigma_{\mathcal{W}}(t) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[W_i(t) - \bar{W}(t)\right]^2}$ and $\sigma_{\mathcal{V}}(t) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left[V_i(t) - \bar{V}(t)\right]^2}$
4. Standardize the observations in the permuted samples:
   $\tilde{\mathcal{W}} = \{\frac{W_i(t) - \bar{W}(t)}{\sigma_{\mathcal{W}}(t)} \mid i = 1, \dots, n\}$ and $\tilde{\mathcal{V}} = \{\frac{V_i(t) - \bar{V}(t)}{\sigma_{\mathcal{V}}(t)} \mid i = 1, \dots, m\}$
5. Calculate $\hat{\tau}_{n,m}$ for the standardized permutation samples $\tilde{\mathcal{W}}$ and $\tilde{\mathcal{V}}$ as described in Section 5 for a chosen measure μ.

The overarching test procedure is again given by the derivation of a permutation distribution and a permutation test decision according to the rules stated in Section 4.

Due to the introduction of the standardization step, the problem described in Section 5.9 is potentially of even greater importance than in the previous sections. As the test relies on standardizing the objects used for calculating the test statistic, it would be of immense importance whether curve fitting is employed and at what step of the procedure it is introduced. For example, if standardization were applied to discrete observations, which are then fitted using a functional basis, the resulting curves would generally not be standardized in the desired way. This can be seen in Figures 12 and 13 in Appendix 11.8 which show samples that were generated by processes with a point-wise mean of 0 and a point-wise variance of 1 and their fitted counterparts. Using curve-fitting could create simulations indicating a strong power against persistence alternatives. However, this power might entirely be based on the artifacts created by the curve fitting procedure.

Therefore, it is advisable to either perform standardization on the fitted curves if curve fitting is desired or to forgo curve fitting. A second option is to rely on comparisons between the discrete observations and the curves created for the approximation of the CvM

type test statistic as described by Bugni and Horowitz 2021 as the variant for discretely observed data. For the purposes of this thesis, the latter approach will be used. However, as the package accompanying this thesis provides functions that can be used to perform the test for samples fitted using a Fourier basis, it would be simple to perform additional simulations comparing the approaches in future extensions of this line of research.

7 Simulation Study

To learn more about the potential practical applications of this method, it is helpful to study its properties in a series of simulations. Bugni and Horowitz 2021 study an array of different setups that give an idea of the performance in a collection of settings, and this thesis will study similar setups to extend the understanding of the method. The simulations presented as part of this thesis have been conducted on *bonna*[2]. *bonna* is the high performance computing cluster provided by the University of Bonn. However, slight modifications of the provided code suffice to run it on personal computers.

All analyses in this thesis were performed with R[3]. As a supplement to this thesis, the two-sample variant of the test taken from Bugni and Horowitz 2021 is implemented in an R package called *PermFDATest*. The R package and all code that has been used to produce the following results are publicly available as part of a GitHub repository[4] that complements this thesis. As previously mentioned in Section 5.9, the CvM type test statistic is calculated by comparing the generated curves at the points where the corresponding discrete observations of each sample curve lie. However, the package additionally provides functions that can be used to perform the test on curves that have been fitted using a Fourier basis. The package makes extensive use of the implementations by Ramsay, Graves, and Hooker 2021 and Wickham et al. 2019.

7.1 Simulation Setup

The simulations presented in this thesis are inspired by the simulation study from Bugni and Horowitz 2021 and shine more light on specific settings that offer insight into the potential benefits and problems of the method. To describe the specific processes used in the simulation, the following section introduces the notation used in Bugni and Horowitz 2021. Deviating from the original paper, these simulations focus on the two-sample variant of the test. Additionally, the data are generated on the closed interval $[0, 1]$ and use a smaller sample size of 20 observations per sample.

[2]https://www.dice.uni-bonn.de/de/hpc/hpc-a-bonn/infrastruktur
[3]R Core Team 2022.
[4]https://github.com/JakobJuergens/Masters_Thesis

Let $\mathcal{T} = \{0, 0.01, 0.02, \ldots, 1\}$ be the index set of the discrete points on the unit interval, then the observations are constructed in the following way.

1. Draw i.i.d. random variables $\{\xi_{i,s}(t) \mid (i,s) \in \{1,\ldots,20\} \times \{1,2\},\ t \in \mathcal{T}\}$ from the standard normal distribution $\mathcal{N}(0,1)$.
2. For all $i = 1,\ldots,20$ and $s = 1,2$, set $\tilde{X}_{i,s}(0) = \xi_{i,s}(0)$.
3. For all $i = 1,\ldots,20,\ s = 1,2$ and $t \in \mathcal{T} \setminus \{0\}$, set $\tilde{X}_{i,s}(t) = \rho_s(t)\tilde{X}_{i,s}(t-0.01) + \xi_{i,s}(t)\sqrt{1-\rho_s^2(t)}$ where $\rho_s(t)$ is a parameter as defined below.
4. For all $i = 1,\ldots,20;\ s = 1,2$ and $t \in \mathcal{T}$, set $X_{i,s}(t) = \mu_s(t) + \sigma_s(t)\tilde{X}_{i,s}(t)$ where $\mu_s(t)$ and $\sigma_s(t)$ are parameters as defined below.

The resulting random variables $\{X_{i,s}(t) \mid t \in \mathcal{T};\ i = 1,\ldots,20;\ s = 1,2\}$ are normally distributed with the following properties.

1. $\mathbb{E}[X_s(t)] = \mu_s(t)$
2. $\mathrm{Var}[X_s(t)] = \sigma_s^2(t)$
3. $\mathrm{Corr}[X_s(t), X_s(t-0.01)] = \rho_s(t)$ $\forall t \in \mathcal{T}$ with $t > 0.01$

This thesis focuses on three distinct violations of the null hypothesis described in the following. For each of the following settings sample 1 is generated by the same process as a reference point to compare the different settings. Therefore, the parameters for sample 1 are chosen as follows.

1. $\mu_1(t) = 0 \quad \forall t \in \mathcal{T}$
2. $\sigma_1(t) = 1 \quad \forall t \in \mathcal{T}$
3. $\rho_1(t) = 0.5 \quad \forall t \in \mathcal{T} \setminus \{0\}$

For the second sample, three violations of the null hypothesis and a benchmark case where the null hypothesis is correct are generated.

1. **Identical Data Generating Processes**
   The first setting is the benchmark, where both samples are generated using the same random function. This setting is meant to support the theoretical points arguing that the permutation tests described in this thesis have the correct size.
2. **Mean Shift**
   The second setting introduces a shifted mean function. Deviating from sample 1, sample 2 has a mean function that is smaller for points in the center of the unit interval.
   $$\mu_2(t) = t(t-1) \quad t \in \mathcal{T}$$
   Pure shifts of the mean function are one scenario where Bugni and Horowitz 2021 describe weak power of the CvM type test.

3. **Correlation Shift**

   The third setting violates the null hypothesis by changing the correlation structure between neighboring observations. Here, observations are less persistent due to a lower but still constant choice of the correlation parameter.

$$\rho_2(t) = 0 \quad \forall t = 0.01, \ldots, 1$$

   For this specific choice, each observation in sample 2 is a sequence of independent observations that follow a standard normal distribution.

4. **Variance Shift**

   In the fourth simulation setting, the point-wise variance function is altered in the following way.

$$\sigma_2(t) = 1 + 0.5t \qquad t \in \mathcal{T}$$

   Therefore, sample 2 exhibits a higher point-wise variance towards the upper end of the unit interval.

Figure 5 shows samples generated by these processes. As can be seen, the difference between the samples is optically quite subdued. Therefore, a high power against these alternatives is not expected. However, any nontrivial power shown in these settings is a good indicator of the qualities of the test.

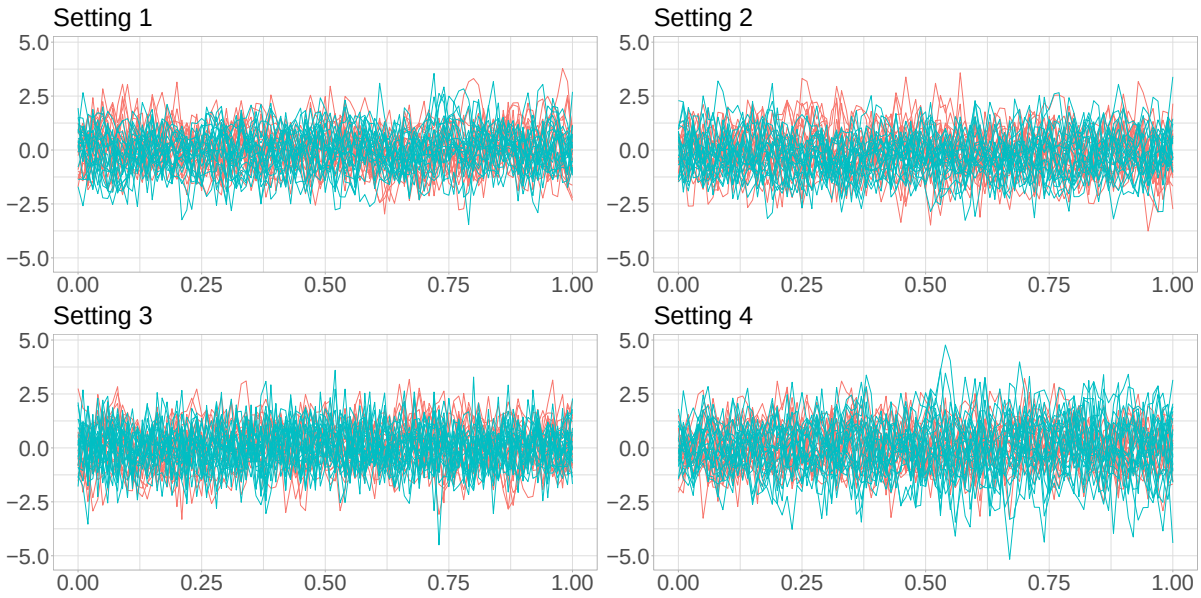

Figure 5: Samples generated for the four settings.
Sample 1 in red, Sample 2 in blue

Before performing the simulations, there are several parameter choices that have to be made, as explained in the theoretical description of the test. These parameter choices have a considerable influence on the properties of the test. Therefore, further simulations comparing different parameter choices might be useful to inform future usage in applied settings.

1. As shown in the mathematical construction of the data, these simulations compare samples consisting of a smaller number of observations per sample $n = 20$ and focus on the two-sample version of the test. Additionally, each observation is constructed from a lower number of discrete observations (101) than in the original paper.

2. Concerning the choice of the truncation parameter K as shown in Equation 30, the authors argue that their simulations showed that the power of the test became flat at about $K = 20$ and chose $K = 25$. For the sake of reproducing their results, the same parameter was used in this thesis.

3. Another parameter that has to be chosen is the number of functions L used in the approximation of the test statistic $\hat{\tau}_{n,m}$ by Monte-Carlo integration as shown in Equation 25. Bugni and Horowitz 2021 chose $L = 4000$ and preliminary testing confirmed that this choice seems reasonable in many settings when combined with reasonable choices for the construction of the measure μ.

4. For the number of permutations Q used for the approximation of the permutation distribution of both test statistics, the authors chose a value of $Q = 500$. This thesis uses the same parameter. However, absent strong computational limitations, a higher choice for Q could benefit the properties of the tests.

5. For the construction of the measure μ the weight function $w(t)$ has to be chosen. This simulation uses the same weight function as in the empirical application in Bugni and Horowitz 2021 as described in Section 5.5.

6. Deviating from the original authors' choice, $(\rho_i)_{i=1,\dots,K}$ is determined by checking the sample standard deviation of the corresponding Fourier coefficients in the first sample obtained by fitting the same basis used in the construction of μ to the observations. These are then multiplied by a constant factor of 2 to increase variability in the functions. This higher variability seemed to improve performance in preliminary tests but should be studied in further simulations.

7. Using the same choice as in the original paper, the simulations use independent standard normal error terms $U_k \quad k = 1, \dots, K$ for the Fourier coefficients shown in Equation 32 for the construction of the measure. As the processes under consideration are Gaussian by construction and thereby have thin tails, this choice should be reasonable following the authors' argumentation.

8. For each setting 1000 simulation runs were performed to approximate the power of the test against the described alternatives. Since some parts of the simulation are computationally quite intensive, this choice was guided by runtime restrictions and Bugni and Horowitz 2021 where the authors also chose to perform 1000 runs per simulation.

7.2 Results

Taking a look at the results from the simulation study, it is interesting to compare the empirical rejection probabilities for different types of violations of the null hypothesis.

Test	$(\alpha_\tau, \alpha_\nu)$	**Setting 1**	**Setting 2**	**Setting 3**	**Setting 4**
Means-Test	$\alpha_\nu = 0.05$	0.058	0.5	0.04709	0.058
CvM-Test	$\alpha_\tau = 0.05$	0.03775	0.55675	0.63978	0.9585
Combined	(0.04, 0.01)	0.04	0.577	0.5977	0.9535
	(0.03, 0.02)	0.043	0.587	0.5516	0.942
	(0.025, 0.025)	0.045	0.576	0.53148	0.938
	(0.02, 0.03)	0.044	0.564	0.51102	0.932
	(0.01, 0.04)	0.046	0.558	0.42929	0.896

Table 1: Empirical Rejection Probabilities

Looking first at setting 1, which is a simulation under the null hypothesis, the rejection probabilities seem compatible with the chosen size of 0.05. However, the size of the rejection frequency for the CvM test seems slightly low, which could be an artifact caused by the relatively low number of simulation runs. A more extensive simulation using 10,000 runs for the CvM type test resulted in a rejection frequency of 0.0469 and therefore seems to confirm this hypothesis. Next, looking at the rejection frequencies of the mean-based test in settings 3 and 4, the test shows only trivial power. This is expected as the samples were explicitly generated by processes sharing a common mean function. For setting 2, which is a mean shift between the processes, the power of both tests is relatively high considering the low sample size of 20. Especially considering the fact that the authors mentioned that the power of the CvM type test in pure mean shift settings is somewhat lacking. However, this can be remedied by the fact that even though the mean shift is rather subdued visually, it is stronger than the mean shift used in the simulation study of Bugni and Horowitz 2021. Due to the change in setting 3, picking up the differences in the persistence structure might be difficult with a measure induced by a random variable truncated at a low truncation parameter K. However, for a choice of $K = 25$, the simulation shows considerable non-trivial power against this alternative. The fact that the combined tests keep a lot of their power as the level of the CvM test decreases while the mean-focused test shows no non-trivial power, indicates that even at higher significance levels, a comparatively high fraction of violations of the null hypothesis can be identified by the CvM test. In setting 4, the rejection frequencies are extremely high, which in light of the other three settings, and the visual presentation of exemplary samples in Figure 5 is unsurprising. In this figure, setting 4 showed the clearest difference between sample 1 and sample 2 and the test was able to pick up on it. Similar to setting 3, the rejection frequencies for the combined tests indicate that the rejection frequency of the CvM type test remained high even at stricter significance levels.

7.3 Simulation - Test for Persistence Alternatives

Using simulations akin to setting 3, it is possible to test the proposed test for persistence alternatives from Section 6. However, due to computational constraints, this simulation currently cannot follow the real test's structure and shall instead serve as a heuristic for the real test's performance.

Primarily, the simulation differs from the proposed test to reduce computational cost while staying reasonably accurate in the proposed simulation setting. The standardization step makes it necessary to perform the comparison between functions that is performed in the calculation of the CvM test statistic in each permutation step. This massively increases the computational cost. Therefore, as a heuristic for the real test, these simulations are performed without a standardization step. However, a data generating process with a mean function $\mu(t) = 0$ and a point-wise variance function $\sigma(t) = 1$ is used. Therefore, in expectation, sample equivalents will share these characteristics. However, individual samples may not. Due to this concession to limited computational resources, the following simulation results could potentially overstate the power of the test against alternatives as described above. This is caused by potential differences in the sample mean and point-wise sample variance present in the permuted samples that would have been removed by standardization. However, in the proposed test structure, the possibility to specifically test for alternatives in the persistence structure is an advantage.

The results for different correlation parameters are given in the following tables. Again, the same reference process with $\rho_1(t) = 0.5$ was used to generate the reference sample.

α_τ	$\rho_2(t) = -0.9$	$\rho_2(t) = -0.5$	$\rho_2(t) = 0$	$\rho_2(t) = 0.5$	$\rho_2(t) = 0.9$
0.05	0.198	0.858	0.640	0.048	0.977
0.01	0.044	0.623	0.393	0.009	0.913
0.001	0.016	0.385	0.197	0.002	0.734

Table 2: Empirical Rejection Probabilities

These results show multiple things. First, as expected, the rejection frequency for the case $\rho_2(t) = 0.5$ is close to the chosen nominal level. Due to the properties of permutation tests, this is not a surprise as it holds by construction.

Second, the other cases, which approximate power against specific alternatives, are interesting as they are in part counter-intuitive. In most settings, one would typically assume that the rejection frequency increases as the parameter that specifies the deviation from the null hypothesis changes away from its equivalent in the reference sample. However, these simulations show that the opposite happens for the case of $\rho_2(t) = -0.9$. Even though the persistence parameter is even further away from 0.5 than for the case of $\rho_2(t) = -0.5$, the rejection frequency decreases. This could be linked to the measure that is used for the calculation of the test statistic. As the parameter decreases, the curves are less persistent,

meaning that Fourier functions of higher frequency would be needed to approximate them. However, if, due to the truncation, it is not possible to use appropriately high-frequent functions in the CvM test, the performance might suffer. This is just a first hypothesis that could be studied in more detail using further simulations with different persistence parameters and different truncation parameters for the basis used to construct the measure. Apart from this unexpected behavior for extremely high-frequent alternatives, the results are as expected, and the test shows considerable non-trivial power against changes in the persistence structure of the process. This is a promising indicator of the potential of this extended testing procedure, and further simulations to study its properties in more detail would be highly interesting due to the test's unique approach.

8 Application

In order to give an example of the real-world merits of the method, this section will compare electricity demand data from Adelaide for different days of the week. This data is provided as part of the *fds*[5] package for R. It consists of half-hourly energy demand in megawatts and was originally used by Magnano and Boland 2007 and Magnano, Boland, and R. J. Hyndman 2008. The data set contains electricity demand curves for 3556 days from 6/7/1997 to 31/3/2007. Some of these curves, specifically a selection observed on Wednesdays and Saturdays fitted using a Fourier basis consisting of 25 functions, is shown in Figure 6.

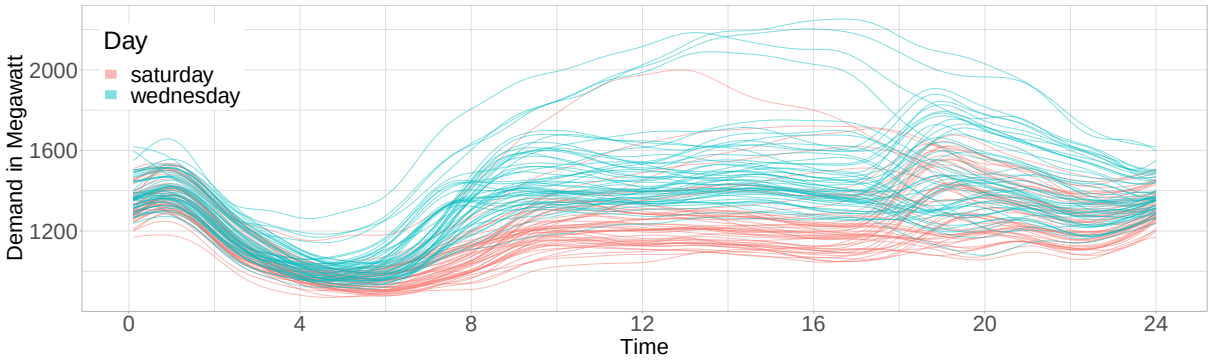

Figure 6: Electricity Demand in Adelaide

The specific question studied with the method presented in this thesis is whether electricity demand on working days and weekends can be seen as if they were generated by the same stochastic process. However, due to information obtained during the data cleaning step that indicates a violation of the i.i.d. assumption, the observations for the weekend will be limited to Saturdays.

[5]Shang and Rob J Hyndman 2018.

8.1 Data Cleaning and Preprocessing

Before actually applying the test from Bugni and Horowitz 2021, it is important to prepare the data for the procedure. As the problem does not have the structure that an experiment as described in Bugni and Horowitz 2021 possesses, some problems must be addressed before using the procedure. These adjustments are described below, and some details are substantiated in Appendix 11.7.

8.1.1 Removal of Mondays and Fridays

One potential problem of this procedure is whether observations on different weekdays or days of the weekend can be seen as independent and identically distributed. While it seems reasonable to assume that demand may be similar on Saturdays and Sundays, it is questionable whether the same can be said for working days. One potential problem is the ramping up and down of industrial production and commercial activity on Mondays and Fridays. Therefore, these days were excluded from the analysis. Thus the weekend is only compared to Tuesdays, Wednesdays, and Thursdays.

8.1.2 Holidays

Furthermore, holidays could appear more frequently on specific weekdays than others. Whereas on weekends, a holiday would not significantly influence the electricity demand due to the already reduced economic activity, this is different for weekdays. Therefore if holidays would occur systematically more often on specific days - such as Thursdays in the case of Germany - this could create problems. Therefore public holidays in Southern Australia, the Australian federal territory containing Adelaide, were excluded from the analysis. A list of holidays that were excluded is given in Appendix 11.7. Additionally, days immediately before and after holidays are excluded for the same reason as Mondays and Fridays. This procedure of eliminating holidays from the data set removed 299 out of 3556 curves from the data set which was used in further steps of the analysis.

8.1.3 Curve Fitting

To employ methods that are suitable to remove trend and seasonal components from the data, it is useful to fit a functional basis to the underlying discrete observations. However, as described in Section 5.9, this can lead to problems if the truncated functional basis is unable to capture the features of the discrete data. However, in the setting explored in this application, visual inspection of the resulting curves showed no signs of artifacts created by the curve fitting process. This makes sense from a theoretical perspective, as the curves are more persistent than those explored in the simulations and do not show pronounced high-frequency variations. Therefore, the truncated basis, consisting of trigonometric functions of lower frequencies, should be able to capture the curves' properties appropriately.

This, in turn, makes the removal of trend and seasonal components from the fitted curves more attractive, as it more closely addresses the functional nature of the underlying data generating processes. Therefore, for the purposes of this simulation, a Fourier basis consisting of 25 functions was fitted to the original data to perform the data cleaning step.

8.1.4 Detrending and Deseasoning

A third potential problem of this data set is its functional time series structure. For example, electricity demand might be systematically higher in the summer months due to the added energy consumption of air-conditioning units. Therefore, a simple interpretation of the data as generated by an i.i.d. process might be unsubstantiated and additional steps have to be made before the procedure can be justified.
Additionally, it might be the case that electricity demand has a trend component that must be removed before this method can reasonably be applied to this data. To combat this low-frequency seasonal component due to the seasons and a potential long-term trend, the following model is used to demean the data as described below. For the estimation of this model, holidays and days immediately before and after holidays are excluded. Mondays and Fridays are included and removed from the data set after the estimation for the creation of the samples used to apply the method described in this thesis.

$$\begin{aligned} f_{demand} =& f_{mean} + f_{trend}(year - 1997) + \sum_{j=2}^{12} \mathbb{1}_{[month = j]} f_{month,j} \\ &+ \sum_{k=2}^{7} \mathbb{1}_{[day = k]} f_{day,k} + f_{random} \end{aligned} \tag{45}$$

The model described in Equation 45 is estimated with the usual theory for function-on-scalar regression, which is described, for example, in Ramsay and Silverman 2005. Specifically, the implementation from the R package *fda*[6] was used to estimate the coefficient functions. Then, the following curves are used for further treatment.

$$\tilde{f} = f_{mean} + \sum_{k=2}^{7} \mathbb{1}_{[day = k]} \hat{f}_{day,k} + \hat{f}_{random} \tag{46}$$

As the resulting estimates of a function-on-scalar regression are functional, it is more convenient to plot the estimates instead of giving the estimated Fourier coefficients. The estimated coefficient functions for the different weekdays are of particular interest as they are directly linked to the problem under consideration. In this case, Sunday is the baseline, and the curves shown in Figure 7 describe the deviation relative to it.

[6] Ramsay, Graves, and Hooker 2021.

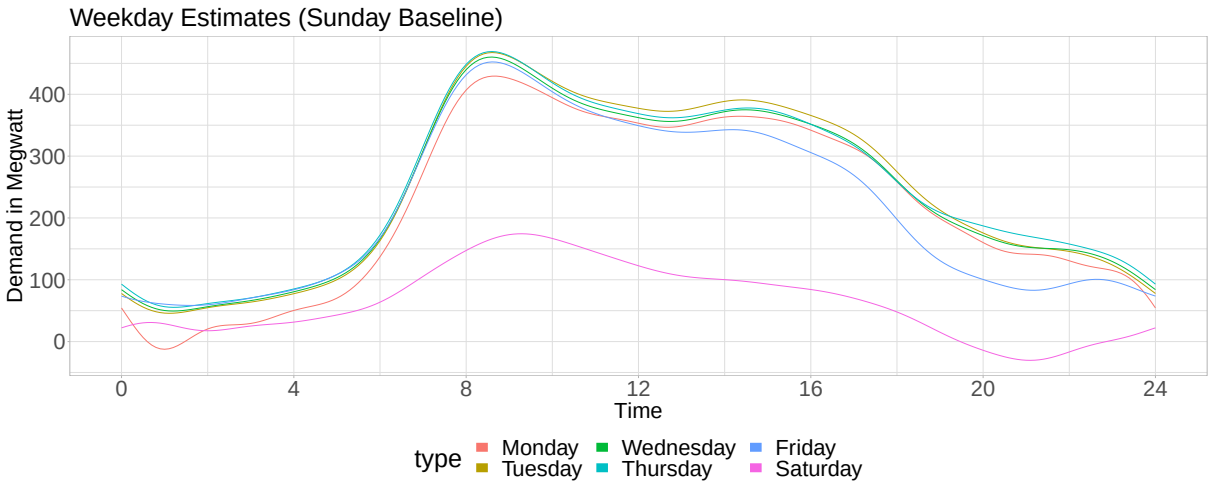

Figure 7: Estimates for the weekdays (Sunday as baseline)

This diagram already hints at a considerable mean shift between the working days and the weekend. As these estimates also hint at a substantial difference between Saturdays and Sundays, treating them as non-identically distributed could be advisable. Therefore, the application is limited to comparing working days and Saturdays as the regression results provide clear evidence against the assumption that Saturday and Sunday can be seen as if the same random variable generated them. Further regression results, including estimates of the constant and the coefficient functions for different months, are shown in Appendix 11.7.2.

8.2 Test from Bugni and Horowitz 2021

After cleaning the data as described in the previous section, the method from Bugni and Horowitz 2021 can now be applied to the cleaned data. To illustrate what these cleaned curves look like, Figure 8 shows randomly chosen curves observed on Wednesdays and Saturdays that were prepared with the procedure described above.

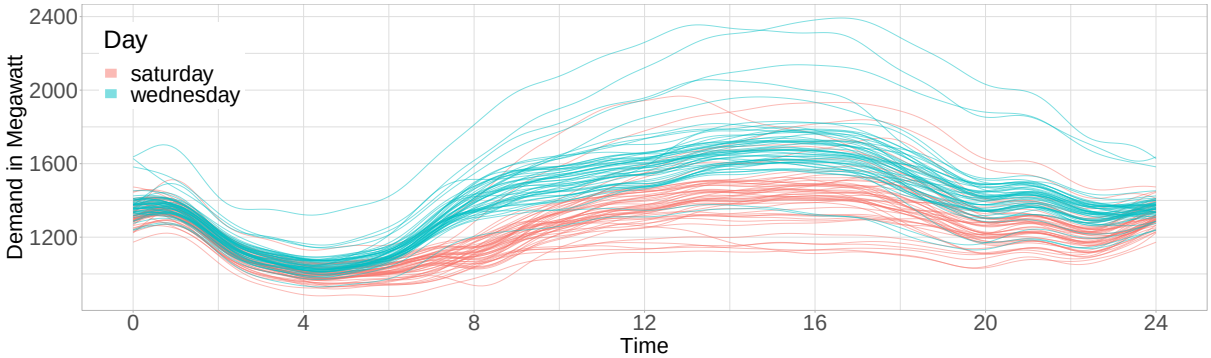

Figure 8: Pre-Processed Electricity Demand in Adelaide

As Figure 8 already shows a considerable mean shift that is more extreme than those shown in the simulation, the test is expected to pick up on these differences. Additionally, the sample size is considerably bigger in this scenario, with 478 observations and 1404 observations corresponding to Saturdays and working days, respectively.

As the power of permutation tests is typically quite sensitive to the sample size, this is another reason to assume strong power against the apparent violation of the null hypothesis. Using the same decision rules as in the simulation study leads to the following rejection decisions.

Test	$(\alpha_\tau, \alpha_\nu)$	**Rejection of the Null?**
Means-Test	$\alpha_\nu = 0.05$	✓
CvM-Test	$\alpha_\tau = 0.05$	✓
Combined	(0.04, 0.01)	✓
	(0.03, 0.02)	✓
	(0.025, 0.025)	✓
	(0.02, 0.03)	✓
	(0.01, 0.04)	✓

Table 3: Results of the Test from Bugni and Horowitz 2021

As mentioned before, these rejections of the null hypothesis are unsurprising when looking at the curves generated for the different days in the data cleaning step. There, it was obvious that the difference in mean alone was stark between workdays and Saturdays. Therefore, it was expected for the mean-based test to find the violation of the null hypothesis in this comparatively large sample problem. Due to the large differences in mean alone, it is also unsurprising that the CvM type test picked up on the violation. This application can therefore serve more as a real-world proof of concept for the test. In most real-world scenarios where the test from Bugni and Horowitz 2021 might be applied a potential violation of the null hypothesis will be less obvious. Especially violations created by differing persistence structures will typically be difficult to identify. Therefore, applying this test to more challenging settings would be interesting. Section 9.4 presents a small selection of settings that might be interesting for further application of the procedures described in this thesis.

9 Outlook

The results presented in this thesis gives an impression of the possibilities of the method developed in Bugni and Horowitz 2021. However, there are problems which came up while writing this thesis that should not go unnamed. Additionally, this outlook presents ideas for further simulations that could improve the understanding of methods such as CvM type tests in a functional data setting and presents real-world scenarios that might be interesting applications for the methods introduced in this thesis.

9.1 Potential Problems

Implementation

One source of problems that might influence the results of the simulations is the chosen implementation. As addressed in Section 5.9, seemingly small choices such as whether to use a functional basis to represent the observations can have a significant impact on the performance and reliability of the simulation. Additionally, better optimization could lead to definite improvements in the usability of the method in real-world applications due to the immense computational costs associated with specific steps of the testing procedure.

Definition of $\mathbb{L}_2[0,1]$

As mentioned multiple times in this thesis, Bugni and Horowitz 2021 use an unconventional definition of $\mathbb{L}_2[0,1]$ that makes some theoretical considerations of the methods considerably harder. Therefore, questions such as how this convention interacts with the type of measure that is used in the CvM type test should be studied further.

9.2 Theoretical Extensions

Variants of the Test based on different Norms

In the classical scalar setting, one common alternative to the CvM test is the Kolmogorov-Smirnov test. Instead of using the squared distance between two objects, it is based on the supremum norm. It could be used analogously to construct an alternative test statistic in the functional setting as shown in Equation 47.

$$\tau_{m,n}^{KS} = \sup_{z \in \mathbb{L}_2[0,1]} |\hat{F}_n(z) - \hat{G}_m(z)| \tag{47}$$

Similar to the approximation by Monte-Carlo integration in the main case, it would have to be approximated in an actual implementation. One obvious approach being the following, where $Z_i \overset{\text{i.i.d.}}{\sim} Z$ and Z is a random function as described in Section 5.5.

$$\hat{\tau}_{m,n}^{KS} = \max_{i=1,\dots,H} \left\{ |\hat{F}_n(Z_i) - \hat{G}_m(Z_i)| \right\} \tag{48}$$

One property of a test such as this is that it would open up the possibility of identifying which function z caused the rejection of the null hypothesis. In practice, this approach might lead to useful information on the difference between the processes that generated the samples.

$$\tilde{\tau} = \mathcal{F}_{distance} \left(\hat{F}_n(z), \hat{G}_m(z) \right) \tag{49}$$

Other variants could be constructed in a similar way from other functional distances between the empirical distribution functions as shown in Equation 49 leading to a plethora of possible test statistics. It would be interesting to compare them both from a theoretical perspective and to use simulations to assess their power in different scenarios.

Extension to Non-separable Processes

As mentioned in Section 5.1, it might be possible to extend the theoretical arguments in favor of the test to processes that are non-separable. This potential extension would be crucial in justifying the application of the test to a broader class of processes. However, some steps in the proofs of Bugni and Horowitz 2021 would have to be approached from a different angle to allow for these generalizations.

Functional Principal Component Version of the CvM type Test

One potential extension of the method could be to use the functional principal components obtained from the samples to construct the measure. This would be interesting both from a theoretical perspective and in practical applications. One aspect that would have to be addressed is that the empirical FPCs do not necessarily form a complete basis of $\mathbb{L}_2[0,1]$. Thus, additional arguments would have to be made to justify the method. However, as the basis used to construct the measure could be better suited to generate functions similar to the original samples, the test's performance might be superior in real-world applications.

9.3 Possible Further Simulations

Comparing different Functional Bases

A functional basis is used for constructing the measure and potentially for fitting smooth curves to discrete observations. Even in settings where observations are not fitted, the basis used to construct the measure heavily influences the test. From a theoretical perspective, the test is only justified for complete orthonormal bases. It would be interesting to study its properties in simulations both for additional complete orthonormal bases and bases that do not have these properties, such as different spline bases.

Comparing choices of $w(t)$ and $(\rho_i)_{i=1,the...,K}$

This thesis only presents simulations for one intuitive choice of $w(t)$ and $(\rho_i)_{i=1,...,K}$. It seems reasonable to assume that these parameter choices could have a significant impact on the performance of the method, and it would therefore be interesting to compare different choices in a structured framework.

Less restrictive Distributions of Fourier Coefficients

The measures constructed as part of the CvM type test described in this thesis drew their Fourier coefficients using independently distributed error terms. Even though this leads to a convenient implementation, using measures induced by random variables generated using error terms with a more general dependence structure might be beneficial to increase the power against specific alternatives. It would be interesting to study the influence on the power of the test in more general settings, both from a theoretical perspective and using simulations.

Simulations for Unbalanced Sample Sizes

The simulations in this thesis only deal with settings where both the reference sample and the second sample contained the same number of curves. Even though there is no obvious theoretical reason for unbalanced sample sizes to influence the power of the test meaningfully, it would be interesting to study settings with unbalanced sample sizes in further simulations.

Comparing Implementations using different orders of Operations

As described in Sections 6.2 and 7.3, the actual implementation could be highly important for the properties of the proposed test for persistence alternatives. This thesis only provides two heuristic simulations to compare the performance of the proposed options due to computational limitations. Thus, further full-scale simulations studying the properties of the proposed procedure are crucial to investigate its potential merits in a real-world setting.

9.4 Potential Applications

As explained in Section 8, the chosen setting is suitable for applying the test, but it is more of a showcase than what this method would be used for in reality. Therefore, it would be interesting to apply the method and the variant for persistence alternatives to data that is more fitting. Two examples that could provide interesting applications for this method are listed in the following.

- **Audio-Curves** and specifically speech recordings are a typical scenario for functional data. The tests described in this thesis could be used to compare samples of recordings of spoken words. One interesting application could be to determine whether a set of recordings was manipulated when there is an appropriate reference sample that is known to be authentic. Especially in computer-generated speech imitation, sometimes known as voice deep-fakes, a test that could reliably distinguish between original and imitated speech recordings could be highly relevant.

- **Movement-Curves**, for example, curves describing the movement of a runner's legs during a 100m sprint might also be an exciting field for potential applications. Given a reference sample of movement curves of runners with a specific alteration to their movement due to, e.g., a sickness, the test might be useful in the context of medical diagnostics.

10 Bibliography

Anderson, T. W. (1962). "On the Distribution of the Two-Sample Cramer-von Mises Criterion". In: *The Annals of Mathematical Statistics* 33.3. Publisher: Institute of Mathematical Statistics, pp. 1148–1159. DOI: 10.1214/aoms/1177704477.

Anderson, T. W. and D. A. Darling (1952). "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes". In: *The Annals of Mathematical Statistics* 23.2, pp. 193–212. DOI: 10.1214/aoms/1177729437.

— (Dec. 1954). "A Test of Goodness of Fit". In: *Journal of the American Statistical Association* 49.268, pp. 765–769. DOI: 10.1080/01621459.1954.10501232.

Bauer, Heinz (2011). *Probability Theory*. De Gruyter. ISBN: 978-3-11-081466-8. DOI: 10.1515/9783110814668.

Boyd, John P. (2006). "Computing the zeros, maxima and inflection points of Chebyshev, Legendre and Fourier series: solving transcendental equations by spectral interpolation and polynomial rootfinding". en. In: *Journal of Engineering Mathematics* 56.3, pp. 203–219. ISSN: 1573-2703. DOI: 10.1007/s10665-006-9087-5.

Bugni, Federico A., Peter Hall, et al. (2009). "Goodness-of-fit tests for functional data". In: *The Econometrics Journal* 12.S1, S1–S18. ISSN: 1368-4221. URL: https://www.jstor.org/stable/23116593.

Bugni, Federico A. and Joel L. Horowitz (2021). "Permutation tests for equality of distributions of functional data". en. In: *Journal of Applied Econometrics* 36.7, pp. 861–877. DOI: 10.1002/jae.2846.

Büning, Herbert and Götz Trenkler (2013). *Nichtparametrische statistische Methoden*. De Gruyter. ISBN: 978-3-11-090299-0. DOI: 10.1515/9783110902990. URL: https://www.degruyter.com/document/doi/10.1515/9783110902990/html?lang=en.

Carleson, Lennart (Jan. 1966). "On convergence and growth of partial sums of Fourier series". In: *Acta Mathematica* 116.none. Publisher: Institut Mittag-Leffler, pp. 135–157. DOI: 10.1007/BF02392815.

Corain, Livio et al. (Sept. 2014). "New insights on permutation approach for hypothesis testing on functional data". In: *Advances in Data Analysis and Classification* 8.3, pp. 339–356. DOI: 10.1007/s11634-013-0162-2.

Cox, Dennis and Jong Lee (2008). "Pointwise Testing with Functional Data Using the Westfall-Young Randomization Method". In: *Biometrika* 95.3, pp. 621–634. ISSN: 0006-3444. URL: https://www.jstor.org/stable/20441490 (visited on 05/26/2022).

Darling, D. A. (1957). "The Kolmogorov-Smirnov, Cramer-von Mises Tests". In: *The Annals of Mathematical Statistics* 28.4, pp. 823–838. DOI: 10.1214/aoms/1177706788.

Dunn, Olive Jean (1961). "Multiple Comparisons among Means". In: *Journal of the American Statistical Association* 56.293. Publisher: Taylor & Francis, pp. 52–64. ISSN: 0162-1459. DOI: 10.1080/01621459.1961.10482090.

Fan, Jianqing and Sheng-Kuei Lin (Sept. 1998). "Test of Significance When Data are Curves". In: *Journal of the American Statistical Association* 93.443, pp. 1007–1021. DOI: 10.1080/01621459.1998.10473763.

Fisz, M. (1960). "On a Result by M. Rosenblatt Concerning the Von Mises-Smirnov Test". In: *The Annals of Mathematical Statistics* 31.2. Publisher: Institute of Mathematical Statistics, pp. 427–429. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177705905.

Gibbons, Jean Dickinson and Subhabrata Chakraborti (2021). *Nonparametric statistical inference*. 6th edition. Boca Raton: CRC Press. ISBN: 978-1-138-08744-6.

Gihman, Iosif Il'ich and Anatoliĭ Vladimirovich Skorokhod (2004). *The Theory of Stochastic Processes I*. Classics in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-20284-4. DOI: 10.1007/978-3-642-61943-4.

Guo, Jia, Bu Zhou, and Jin-Ting Zhang (Aug. 2018). "Testing the equality of several covariance functions for functional data: A supremum-norm based test". en. In: *Computational Statistics & Data Analysis* 124, pp. 15–26. DOI: 10.1016/j.csda.2018.02.002.

— (July 2019). "New Tests for Equality of Several Covariance Functions for Functional Data". In: *Journal of the American Statistical Association* 114.527. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2018.1483827, pp. 1251–1263. DOI: 10.1080/01621459.2018.1483827.

Hall, Peter and Nader Tajvidi (June 2002). "Permutation tests for equality of distributions in high-dimensional settings". In: *Biometrika* 89.2, pp. 359–374. DOI: 10.1093/biomet/89.2.359. (Visited on 05/26/2022).

Hall, Peter and Ingrid Van Keilegom (2007). "TWO-SAMPLE TESTS IN FUNCTIONAL DATA ANALYSIS STARTING FROM DISCRETE DATA". In: *Statistica Sinica* 17.4, pp. 1511–1531. ISSN: 1017-0405. URL: https://www.jstor.org/stable/24307686.

Hoeffding, Wassily (June 1952). "The Large-Sample Power of Tests Based on Permutations of Observations". In: *The Annals of Mathematical Statistics* 23.2, pp. 169–192. DOI: 10.1214/aoms/1177729436.

Hsing, Tailen and Randall L. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley series in probability and statistics. John Wiley and Sons, Inc. ISBN: 978-0-470-01691-6.

Jeribi, Aref (Jan. 2021). *Perturbation Theory for Linear Operators*. Springer Singapore Pte. Limited. ISBN: 9789811625275.

Kokoszka, Piotr and Matthew Reimherr (2021). *Introduction to functional data analysis*. First issued in paperback. Texts in statistical science series. CRC Press. ISBN: 978-1-03-209659-9 978-1-4987-4634-2.

Lee, Jong, Dennis Cox, and Michele Follen (Mar. 2015). "A Two Sample Test for Functional Data". In: *Communications for Statistical Applications and Methods* 22, pp. 121–135. DOI: 10.5351/CSAM.2015.22.2.121.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. en. Springer Texts in Statistics. Springer New York. ISBN: 978-0-387-98864-1 978-0-387-27605-2. DOI: 10.1007/0-387-27605-X.

Magnano, L. and J. W. Boland (Nov. 2007). "Generation of synthetic sequences of electricity demand: Application in South Australia". en. In: *Energy* 32.11, pp. 2230–2243. DOI: 10.1016/j.energy.2007.04.001.

Magnano, L., J. W. Boland, and R. J. Hyndman (2008). "Generation of synthetic sequences of half-hourly temperature". en. In: *Environmetrics* 19.8, pp. 818–835. DOI: 10.1002/env.905.

Pettitt, A. N. (1976). "A Two-Sample Anderson–Darling Rank Statistic". In: *Biometrika* 63.1, pp. 161–168. DOI: 10.2307/2335097.

Pomann, Gina-Maria, Ana-Maria Staicu, and Sujit Ghosh (2016). "A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 65.3, pp. 395–414. ISSN: 0035-9254. URL: https://www.jstor.org/stable/24773028.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ramsay, J. O., Spencer Graves, and Giles Hooker (2021). *fda: Functional Data Analysis*. R package version 5.5.1. URL: https://CRAN.R-project.org/package=fda.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer New York. ISBN: 978-0-387-40080-8 978-0-387-22751-1. DOI: 10.1007/b98888.

Richmond, Tom (July 2020). *General Topology: An Introduction*. en. De Gruyter. DOI: 10.1515/9783110686579.

Rosenblatt, M. (1952). "Limit Theorems Associated with Variants of the Von Mises Statistic". In: *The Annals of Mathematical Statistics* 23.4. Publisher: Institute of Mathematical Statistics, pp. 617–623. ISSN: 0003-4851. URL: https://www.jstor.org/stable/2236587.

Shang, Han Lin and Rob J Hyndman (2018). *fds: Functional Data Sets*. R package version 1.8. URL: https://CRAN.R-project.org/package=fds.

Shonkwiler, Ronald W. and Franklin Mendivil (Aug. 2009). *Explorations in Monte Carlo Methods*. en. Springer Science & Business Media. ISBN: 978-0-387-87836-2.

Skorohod, A. V. (1974). *Integration in Hilbert Space*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-65634-7. DOI: 10.1007/978-3-642-65632-3.

Westfall, Peter H. and S. Stanley Young (1993). *Resampling-based multiple testing: examples and methods for P-value adjustment*. Wiley series in probability and mathematical statistics. New York: Wiley. ISBN: 978-0-471-55761-6.

Wickham, Hadley et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

11 Appendix

11.1 Informal Intuition

For potential readers unfamiliar with functional data analysis and permutation testing, the initial hurdle of reading this thesis without any intuitive understanding might be high. Therefore, this short informal introduction should serve as a primer that gives an intuitive idea about the following questions.

1. What setting are we in?
2. What do we want to test?
3. How do we want to test it?
4. Why should I care?

By choice, this section will **not** be formal and it will **not** be precise so as not to introduce too much detail that could hinder an intuitive understanding.

Answering question one is relatively simple. This thesis deals with a statistical problem where observations are continuously observed curves. But what does that mean intuitively? One scenario that is easy to imagine is a curve over time for some variable such as speed: When driving a car at the 24-hour race at Le-Mans, it has a speed at every point in time during the race - for example, 155.3 km/h at 2 hours, 33 minutes and 7 seconds after the race has started. Observations in the sense of this thesis are similar to this curve: at every value of x in some compact interval $[a, b] \subset \mathbb{R}$, we observe a value y.

Question two is also simple. We want to test if two samples of curves are similar in a specific way. As in many statistical problems, we interpret the observations as realizations of some random variable. The only difference is that these observations are curves. So we can ask the question of whether the curves in the two samples are dissimilar enough for us to say confidently that they were not generated by the same random variable. Staying with the example of Le-Mans, we could have two identical cars with two equally skillful drivers collect 24-hour speed curves for the next ten years. We do this to determine if two different types of fuel change the way the cars act on the track. We now have two samples, each containing 3650 speed curves in this hypothetical scenario. We want to test statistically whether changing the fuel made any difference.

Question three is more complicated. Therefore, let's give the correct intuition by answering a more straightforward question instead. Let us return to the scalar setting for a moment - each observation is just a real number $x \in \mathbb{R}$ again. We want to determine whether two samples are different enough to confidently say: "These samples are too dissimilar to be generated by the same random variable." If you have some statistical knowledge, your first intuition might be to perform a Kolmogoroff-Smirnov, CvM, or Anderson-Durbin test. However, let us take one step back first and think about a more general idea.

From an intuitive point of view, if the same random variable generated the two samples, the effect of randomly switching observations within each sample and across samples on an appropriately chosen summary statistic should be small. Formalizing this idea leads to the concept of permutation tests. By randomly permuting the samples and calculating a chosen test statistic on the permuted samples, one could derive a distribution for the test statistic that can be used for testing. If the test statistic calculated on the non-permuted samples is comparatively extreme, this could be an indicator that the samples were not generated by the same random variable.

Why you should care about this is the most challenging question. I care about it because it is cool from a mathematical perspective. Nevertheless, if you care about real-life problems, there are also good reasons to be interested. Economic data is being observed at increasing frequencies as technology improves. Many problems that were previously low dimensional due to data constraints or suitable for classical time series methods are becoming more complex as methodological challenges such as high-dimensionality and extreme correlation of neighboring observations arise. Functional data analysis is a comparatively new approach to many of these problems that transform some of these problems into strengths by acknowledging the functional structure of the underlying processes. Permutation tests are also comparatively new, at least in our ability to apply them on a larger scale.

11.2 $\mathbb{L}_2[0,1]$ as defined in Bugni and Horowitz 2021

To distinguish between the typical case presented in the previous parts of this thesis, let $\mathbb{L}_2[0,1]$ denote the Hilbert space of square-integrable functions and $\mathbb{L}_2^*[0,1]$ the square-integrable functions under the the convention from Bugni and Horowitz 2021. Using $\mathbb{L}_2^*[0.1]$ creates some interesting theoretical challenges, as the resulting object is in fact not a Hilbert space. To understand the theoretical problems that can occur, it is necessary to first introduce some additional concepts to illustrate the challenges.

Definition 11.1 (Norm and Seminorm)
A function $p : \mathbb{V} \to \mathbb{F}$ on a vector space $\mathbb{V}$ over a field $\mathbb{F}$ is called a norm if the following four conditions hold for all $v, u \in \mathbb{V}$ and $\lambda \in \mathbb{F}$.

1. $p(v+u) \leq p(v) + p(u)$
2. $p(\lambda v) = |\lambda| p(v)$
3. $p(v) \geq 0$
4. $p(v) = 0 \implies v = 0$

If $p : \mathbb{V} \to \mathbb{F}$ fulfills only properties (1) to (3) it is called a seminorm.

In the same way a norm induces a distance on its corresponding normed vectorspace, a seminorm p induces a so-called pseudometric d. It is given by $d(v,u) = p(u-v)$. In the

same way a vector space together with a metric induced by a norm is a metric space, a seminorm together with an associated space leads to the concept of pseudometric spaces.

Definition 11.2 (Pseudometric Space)
A pseudometric space (X, d) is a set X together with a function $d : X \times X \to \mathbb{R}_{\geq 0}$, such that $\forall x, y, z \in X$ the following properties hold.

1. $d(x, x) = 0$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Therefore, deviating from a metric space, two distinct points in a pseudometric space can have a distance of zero $d(x, y) = 0$ for $x \neq y$.

That $\mathbb{L}_2^*[0, 1]$ is not a Hilbert space becomes clear, when checking for the properties of the norm induced by the inner product $\|v\| = \sqrt{\langle v, v \rangle}$. One of the properties that has to be fulfilled by a norm is $\|v\| = 0 \iff v = 0$. Let $f : [0, 1] \to \mathbb{R}$ be given by $f(x) = \mathbb{1}[x = 0.5]$. Then evaluation of the following expression creates a contradiction to the norm properties.

$$\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\int_0^1 [f(t)]^2 \, \mathrm{d}t} = 0 \tag{50}$$

As f is not the zero element of this space, this is a violation of positive definiteness. Positive definiteness applied to the case at hand, states that $\forall v \in \mathbb{L}_2[0, 1] \quad \|v\| = 0 \implies v(x) = 0 \quad \forall x \in [0, 1]$. Instead, $\|v\| = \sqrt{\langle v, v \rangle}$ is a seminorm and the defined space should more correctly be treated as a pseudometric space. In this context, it is useful to consider the concept of Hausdorff spaces. Richmond 2020 define Hausdorff spaces as follows.

Definition 11.3 (Hausdorff Space)
A topological space is called a Hausdorff space if $\forall a \neq b$, there exist a neighborhood N_a of a and a neighborhood N_b of b such that $N_a \cap N_b = \emptyset$. This property is also called neighborhood-separability.

One problem of $\mathbb{L}_2^*[0, 1]$ is that the space with the topology induced by the obvious seminorm would not be Hausdorff. Thus, limits in the later part of Bugni and Horowitz 2021 would not be defined. The second problem is that it is not clear how a Schauder basis would be defined for a pseudometric space such as $\mathbb{L}_2^*[0, 1]$ and that typical existence results for orthonormal bases might not be available. For example, the Fourier basis is an almost everywhere basis of $\mathbb{L}_2[0, 1]$ as shown by Carleson 1966. However, for many square-integrable functions, point-wise convergence of a weighted sum of basis functions, as shown in Equation 29, is not fulfilled. This is simple to see for any function that does not have identical values on both sides of its domain. In more basic terms, this implies that using the Fourier basis, it is impossible to construct some functions in $\mathbb{L}_2^*[0, 1]$ even using a non-truncated representation as shown in Equation 29.

This, in turn, implies that a probability measure that is constructed as shown in Section 5.5 cannot give positive weight to many functions in the space $\mathbb{L}_2^*[0,1]$. Namely, for any function in $\mathbb{L}_2[0,1]$ for which the Fourier series fails to converge point-wise, it is impossible to assign a positive probability. However, as the null hypothesis is formulated with respect to μ, this does not formally create any problems as these functions are irrelevant in the formulated null hypothesis. In an application, this should, however, be kept in mind as the construction of the test might lead to unintuitive behavior in less controlled settings.

11.3 Asymptotic Distribution of the Cramér-von Mises Test

As shown by Rosenblatt 1952 and Fisz 1960, under the null hypothesis that both samples were independently generated by random variables sharing the same distribution function, it is possible to derive the following limiting distribution of $C_{m,n}$.

$$\begin{aligned} C_{m,n} &\xrightarrow{\mathrm{d}} \int_0^1 \left(Z(u) + (1+\lambda)^{-\frac{1}{2}} f(u) - \left[\frac{\lambda}{1+\lambda}\right]^{\frac{1}{2}} g(u) \right)^2 \mathrm{d}u \\ &\text{as} \quad n \to \infty, \quad m \to \infty, \quad \frac{n}{m} \to \lambda \in \mathbb{R} \end{aligned} \tag{51}$$

Here, $f(u)$ and $g(u)$ are the probability density functions of the corresponding random variables and $Z(u)$ is a Gaussian stochastic process with the following properties.

- $\mathbb{E}\left[Z(u)\right] = 0 \quad \forall u \in [0,1]$
- $Cov\left(Z(u), Z(v)\right) = \min(u,v) - uv \quad \forall u, v \in [0,1]$

11.4 Proof from Hoeffding 1952

Theorem as stated in Lehmann and Romano 2005

Suppose X^N has distribution P_N in $\mathcal{X}_N$, and G_N is a finite group of transformations from $\mathcal{X}_N$ to $\mathcal{X}_N$. Let $\mathcal{G}_N$ be a random variable that is uniform on G_N. Also let $\mathcal{G}'_N$ have the same distribution as $\mathcal{G}_N$, with X^N, $\mathcal{G}_N$, and $\mathcal{G}'_N$ mutually independent. Suppose, under P_N,

$$\left(T_N(\mathcal{G}_N X^N), T_N(\mathcal{G}'_N X^N)\right) \to_d (T, T'), \tag{52}$$

where T and T' are independent, each with common cumulative distribution function $R(t)$. Then, under P_N,

$$\hat{R}_N(t) \to_P R(t) \tag{53}$$

for every t which is a continuity point of $R(t)$.

Proof as presented in Lehmann and Romano 2005

Let t be a continuity point of $R(t)$. Then,

$$\mathbb{E}_{P_N}\left[\hat{R}_N(t)\right] = P_N\left[T_N(\mathcal{G}_N X^N) \leq t\right] \to R(t), \tag{54}$$

by the convergence hypothesis shown in Equation 52. It therefore suffices to show that $Var_{P_N}\left[\hat{R}_N(t)\right] \to 0$ or, equivalently, that

$$\mathbb{E}_{P_N}\left[\hat{R}_N^2(t)\right] \to R^2(t). \tag{55}$$

But,

$$\begin{aligned}\mathbb{E}_{P_N}\left[\hat{R}_N^2(t)\right] &= \frac{1}{M_N^2}\sum_{g\in G_N}\sum_{g'\in G_N} P_N\left[T_N(gX^N) \le t \ \wedge \ T_N(g'X^N) \le t\right] \\ &= P_N\left[T_N(\mathcal{G}_N X^N) \le t \ \wedge \ T_N(\mathcal{G}'_N X^N) \le t\right] \to R^2(t),\end{aligned} \tag{56}$$

again by the convergence hypothesis. Hence $\hat{R}_N(t) \to R(t)$ in P_N probability.

11.5 Argument for Finiteness

The following argument can be made concerning the finiteness of the mean.

$$\begin{aligned}&\hat{F}_n(z) \in [0,1] \wedge \hat{G}_m(z) \in [0,1] \implies \left[\hat{F}_n(z) - \hat{G}_m(z)\right]^2 \in [0,1] \quad \forall z \in \mathbb{L}_2[0,1] \\ &\int_{\mathbb{L}_2[0,1]} 1 \ \mathrm{d}\mu(z) = 1 \wedge \int_{\mathbb{L}_2[0,1]} 0 \ \mathrm{d}\mu(z) = 0 \implies \int_{\mathbb{L}_2[0,1]} \left[\hat{F}_n(z) - \hat{G}_m(z)\right]^2 \mathrm{d}\mu(z) \in [0,1]\end{aligned} \tag{57}$$

11.6 Multiple Testing

When testing statistical hypotheses, it is often helpful or even necessary to test multiple hypotheses independently of each other. One setting where this could be useful is when trying to combine the desirable properties of two tests, as done by Bugni and Horowitz 2021. If the tests do not perfectly depend on each other, this creates a problem relating to the size of the combined test.

Definition 11.4 (Family-wise Error Rate)
The family-wise error rate is the probability of making at least one type-1 error when performing multiple hypothesis tests.

The most straightforward correction for this multiple testing problem is the so-called Bonferroni Correction. Introduced by Dunn 1961, it is based on Boole's Inequality shown in Equation 58, which is sometimes referred to as the Bonferroni Inequality.

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] \le \sum_{i=1}^{\infty} \mathbb{P}\left[A_i\right] \tag{58}$$

for a countable set of events $A_1, A_2, \ldots$. This inequality motivates the idea to construct a combined test by rejecting the null hypothesis if any individual test indicates a violation of the null hypothesis. It is then necessary to control the family wise error rate when

performing multiple tests by choosing the individual tests' size such that they add up to the desired overall family-wise error rate. However, as tests might be correlated, this could lead to an overall test that is conservative. Still, if little information is known about the correlation structure of the individual test statistics, this idea ensures that the family-wise error rate is at most the sum of the individual test sizes.

11.7 Application Data Cleaning

11.7.1 Excluded Holidays

Holidays that were excluded in the analysis are the following[7]: New Year's Day, Australia Day, March Public Holiday, Good Friday, Holy Saturday, Easter Monday, Anzac Day, Queen's Birthday, Labour Day, Christmas Day, Christmas Eve, Christmas Day, Proclamation Day and New Year's Eve. Sunday is nominally a public holiday in South Australia. Easter Sunday is therefore not a special public holiday, but due to its prominence it was excluded in this analysis.
As Australia replaces some holidays that fall on weekends with substitute holidays on the next working day, these were also excluded. For South Australia, this can occur for New Year's Day, Australia Day, ANZAC Day, Christmas Day, Proclamation Day and New Year's Eve.

11.7.2 Detrending and Deseasoning Results

Figures 9 and 10 show the estimates for the coefficient functions as specified in Equation 45. These curves were subtracted from the respective observations to eliminate the influence of, for example, the month the data was observed.

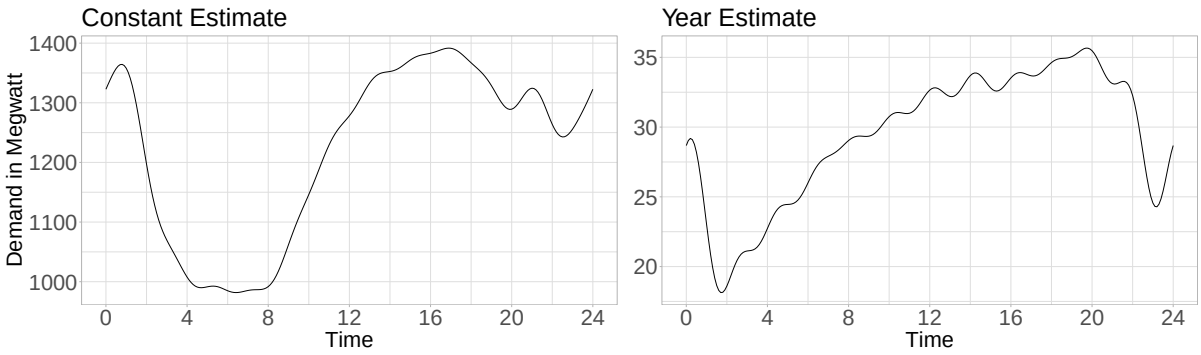

Figure 9: Estimates for the constant and year

[7]These were taken from `https://www.australia.gov.au/public-holidays` accessed on 20.05.2022.

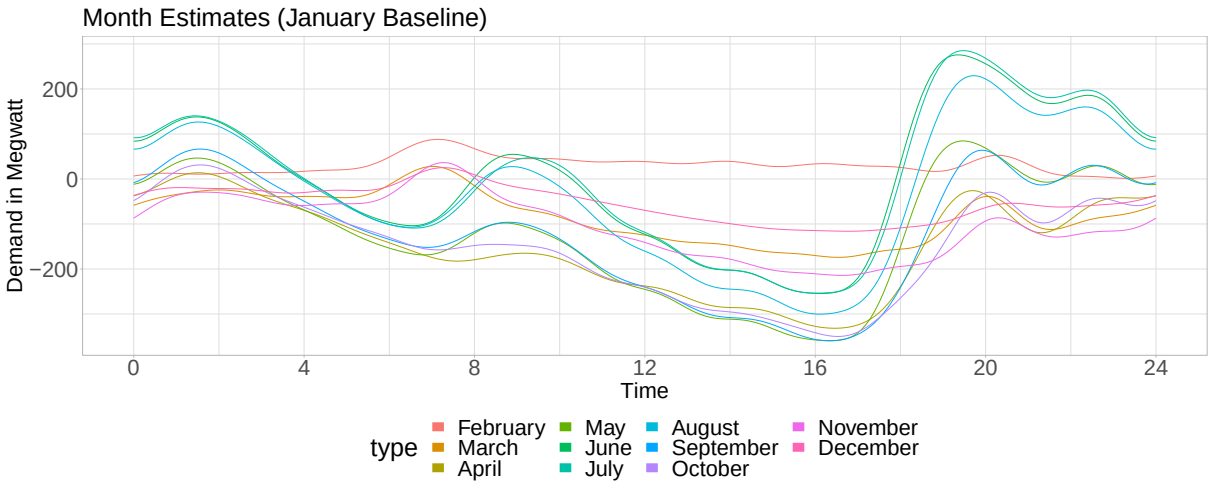

Figure 10: Estimates for the months (January as baseline)

11.8 Additional Figures

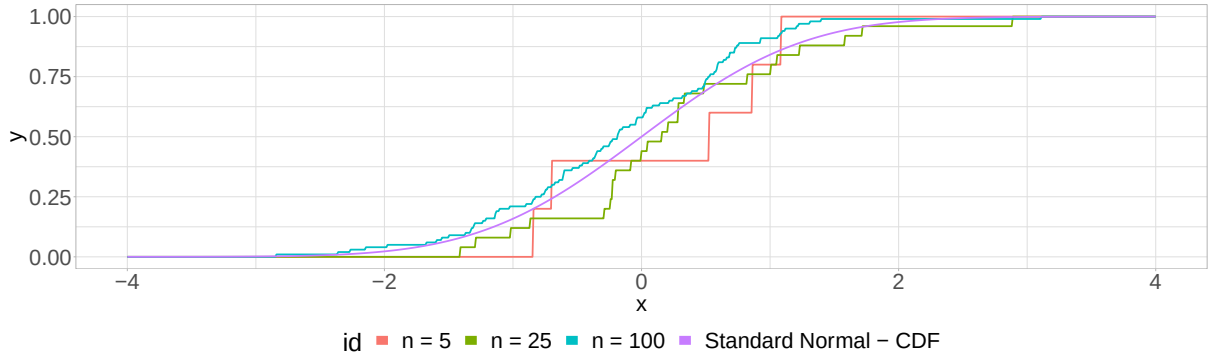

Figure 11: Empirical Distribution Functions for different samples drawn from a Standard Normal Distribution

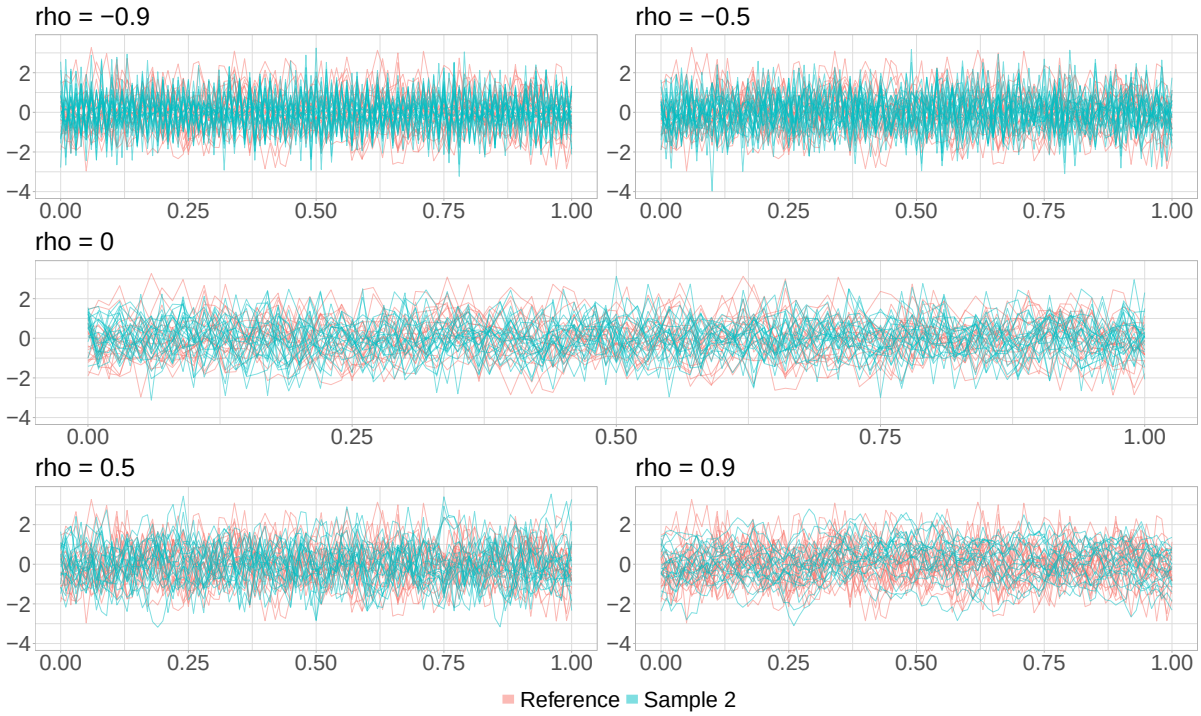

Figure 12: Original Samples for Persistence Test

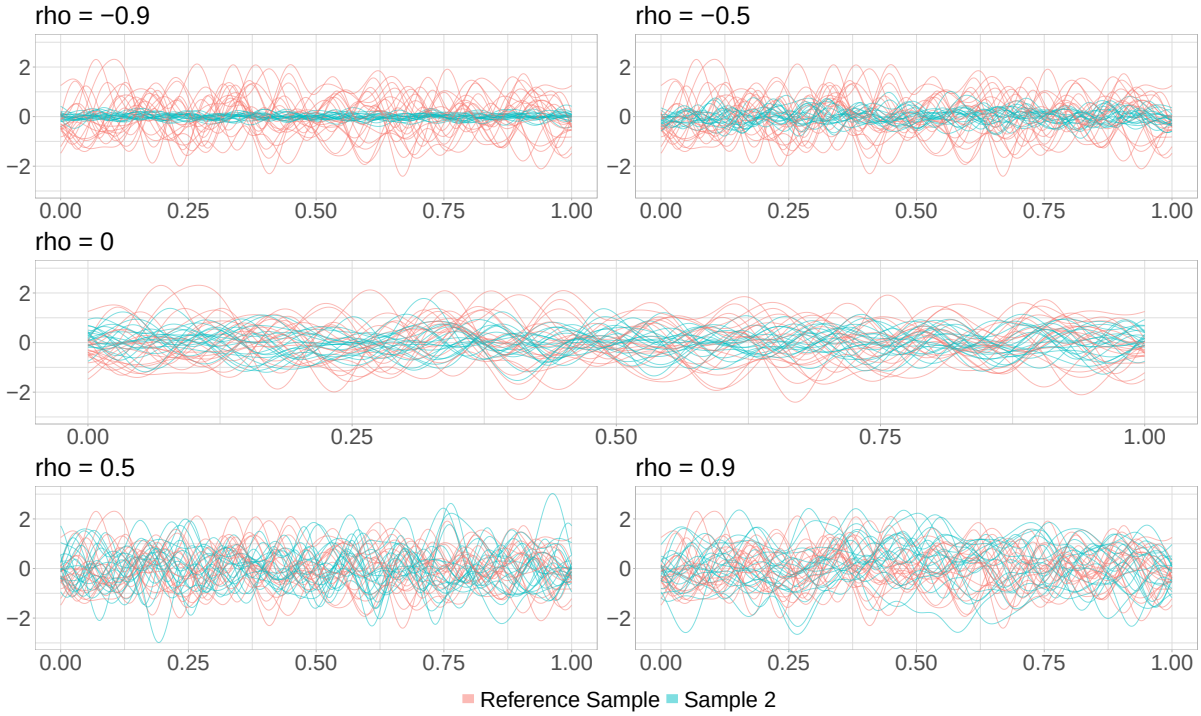

Figure 13: Samples for Persistence Test fitted using a Fourier Basis with 25 functions

Versicherung an Eides statt

Ich versichere hiermit, dass ich die vorstehende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass die vorgelegte Arbeit noch an keiner anderen Hochschule zur Prüfung vorgelegt wurde und dass sie weder ganz noch in Teilen bereits veröffentlicht wurde. Wörtliche Zitate und Stellen, die anderen Werken dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall kenntlich gemacht.

Bonn, 10.07.2021 _____

Jakob R. Juergens