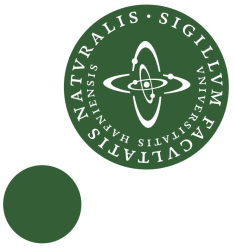
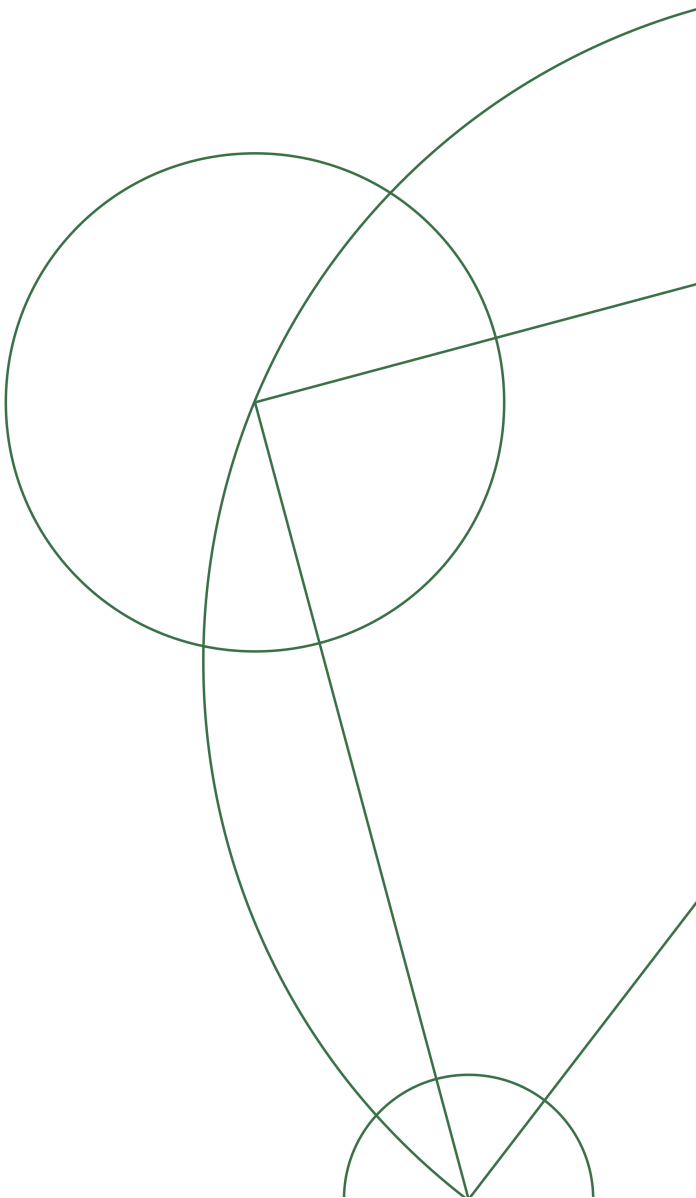
Department of Mathematical Sciences
UNIVERSITY OF COPENHAGEN



Anton Rask Lundborg Hansen

Conditional Independence Testing in Hilbert Spaces



Master's thesis in statistics
Department of Mathematical Sciences
University of Copenhagen

Advisor:
Jonas Peters

July 2019

Abstract

In this thesis we investigate conditional independence testing in Hilbert spaces with a particular focus on infinite-dimensional and separable spaces. We review Shah & Peters' construction of the Generalised Covariance Measure (GCM) by providing extra details on the proof of pointwise asymptotic level and of uniform asymptotic level. We also show that the GCM has pointwise asymptotic level when testing $X \perp\!\!\!\perp Y \mid Z$ for X and Y univariate real-valued random variables and Z a functional random variable when the relationship between X and Z and Y and Z can be explained using a scalar-on-function linear regression model. We then proceed to generalise the GCM to separable Hilbert spaces (possibly of infinite dimension), thus constructing the Generalised Hilbert Space Covariance Measure (GHSCM) and prove that the GHSCM has pointwise asymptotic level. We show that the GHSCM has pointwise asymptotic level when testing $X \perp\!\!\!\perp Y \mid Z$ for X, Y and Z functional random variables when the relationship between X and Z and Y and Z can be explained using a function-on-function linear regression model. Finally we verify the results in a simulation study where we show that there exists cases where the GHSCM is better at detecting conditional independence than the GCM.

CONTENTS

Introduction

1.1 Motivation and overview

In this thesis we consider the problem of conditional independence testing for random variables with values in a Hilbert space. The primary interest is in the separable and infinite-dimensional case but the theory will also apply for finite-dimensional Hilbert spaces. This problem is not merely of theoretical interest but recent interest in the functional data analysis paradigm has made the study of infinite-dimensional random variables all the more relevant. In this framework, instead of observing n i.i.d. samples of some real-valued random variables, we observe n curves, that are generated in some way from discrete data (typically through some form of smoothing). These curves can be viewed as elements of $L^2[0,1]$ or $C[0,1]$ depending on the context and as thus techniques for working with random variables in infinite-dimensional spaces are needed.

The interest in developing conditional independence tests has also increased in the last few decades for a variety of reasons. Conditional independence relations are the fundamental components of graphical models, that have become applied increasingly often in the realm of computational statistics [15]. In the field of causal inference too, the language of conditional independence is often found and applied in many foundational algorithms such as the PC algorithm and also in more modern methods such as invariant prediction. [19]. One could hope that a conditional independence test for random variables in an infinite-dimensional Hilbert space would allow for new developments in causal inference for functional data.

1.2 Outline

The starting point for this thesis is a test of conditional independence for univariate real-valued random variables; the Generalised Covariance Measure (GCM) as constructed by

Shah & Peters in [25]. We will repeat the development of this test in Chapter 2 and also introduce some preliminaries from probability theory and test theory. We will also provide more details on the uniform results given in the original article. The chapter ends with an explicit conditional independence test when X and Y are univariate real-valued random variables and Z is a functional variable in $L^2[0,1]$ and where the relationship between X and Z and Y and Z is assumed to be linear. In Chapter 3 we describe the various properties of Hilbert spaces including linear functionals and operators on and between Hilbert spaces. Furthermore we develop a theory of integration for separable Hilbert spaces. In Chapter 4 we develop a framework for probability theory on separable Hilbert spaces including how to define a random variable on such a space and how to calculate the mean and covariance of such a random variable. We also touch upon how to define conditional expectations on Hilbert spaces and give a brief idea of how to define linear models on Hilbert spaces. In Chapter 5 we generalize the GCM to the Hilbertian case and prove that test has pointwise asymptotic level. We show that we can use the GHSCM to construct a conditional independence test with pointwise asymptotic level when X, Y and Z are functional random variables and the relationship between X and Z and Y and Z is linear. Finally a small simulation study is conducted where the GCM is compared to the GHSCM.

1.3 Contributions

Below is a list of the most significant contributions to the literature:

- Providing further analysis on the proof of uniform asymptotic level for the GCM.
- A novel application of the GCM to the case of testing $X \perp\!\!\!\perp Y \mid Z$ when X and Y are univariate real-valued and Z is functional.
- A self-contained introduction to Bochner integration and random variables on Hilbert spaces.
- An extension of the GCM to infinite-dimensional Hilbert spaces and a proof that the extensions holds pointwise asymptotic level.
- A novel construction of a test of $X \perp\!\!\!\perp Y \mid Z$ when X, Y and Z are functional.

Preliminaries, test theory and the Generalised Covariance Measure

In this chapter we give a self-contained description of the Generalised Covariance Measure for univariate real-valued random variables as originally constructed by Shah & Peters [25].

2.1 Probabilistic preliminaries

We begin by first restating some preliminary results, focusing on the definitions of independence and conditional independence. σ-algebras turn out to be a convenient language to express these notions in great generality. The appendix contains a summary of measure-theoretic probability for the uninitiated. Most of the theory and development follows [9] and [27].

In the following definitions we will concentrate on independence and conditional independence of two σ-algebras or two random variables for brevity but the notions could be expanded to also include countable or uncountable families of σ-algebras or random variables.

Definition 2.1.1 (Independence of σ-algebras)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{F}_1$ and $\mathbb{F}_2$ be sub-σ-algebras of $\mathbb{F}$. If

$$P(F_1 \cap F_2) = P(F_1)P(F_2), \quad \forall F_1 \in \mathbb{F}_1, F_2 \in \mathbb{F}_2,$$

we say that $\mathbb{F}_1$ is *independent* of $\mathbb{F}_2$ and write $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2$.

In practice we will always work with random variables but independence of random variables is defined through independence of σ-algebras as we shall see. Recall that for a random variable X defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in the measurable space $\mathcal{X}, \mathbb{E})$, we define $\sigma(X)$ to be the smallest sub-σ-algebra of $\mathbb{F}$, that makes X $\mathbb{F}-\mathbb{E}$-measurable, i.e. for all $E \in \mathbb{E}$ we have $X^{-1}(E) \in \sigma(X)$. We can write this set explicitly as $\sigma(X) = \{X^{-1}(E) \mid E \in \mathbb{E}\}$ or in more probabilistic language we can say that $\sigma(X)$ contains all sets of the form $(X \in E)$ for $E \in \mathbb{E}$.

Definition 2.1.2 (Independence of random variables)**.** Let X and Y be random variables defined on the same probability space $(\Omega, \mathbb{F}, P)$ with values in the measurable spaces $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{G})$ respectively. We say that the random variables X and Y are *independent* if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent and we write $X \perp\!\!\!\perp Y$.

Remark 2.1.3 (Equivalence of independence definitions)**.** Note that independence of the σ-algebras generated by two random variables X and Y with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{G})$ respectively can be written explicitly as

$$P(X \in E, Y \in G) = P(X \in E)P(Y \in G), \quad \forall E \in \mathbb{E}, G \in \mathbb{G},$$

which is the elementary definition of independence of random variables.

The two definitions above are compatible in the sense that given two independent random variables X and Y by definition the σ-algebras generated by the variables are independent. If instead we are given two independent σ-algebras $\mathbb{F}_1$ and $\mathbb{F}_2$ and construct two random variables X and Y into two possibly different measurable spaces such that X is $\mathbb{F}_1$-measurable and Y is $\mathbb{F}_2$-measurable, then it is straightforward to see that $X \perp\!\!\!\perp Y$.

The intuition for independence of random variables is that independent random variables do not affect each other. If $X \perp\!\!\!\perp Y$ then knowing something about X does not tell me anything about Y. We will not go over every property of independent random variable here but we will note following essential characterization that we are going to use later.

Theorem 2.1.4 (Characterization of independence)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and X and Y be random variables into the measurable spaces $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{G})$ respectively. Then for all $f : \mathcal{X} \to \mathbb{R}$, $g : \mathcal{Y} \to \mathbb{R}$ Borel measurable and bounded functions we have

$$E(f(X)g(Y)) = E(f(X))E(g(Y))$$

if and only if $X \perp\!\!\!\perp Y$.

Proof.
It is easy to see that if X and Y are independent and f and g are measurable then $f(X)$ and $g(Y)$ are independent. By the boundedness of f and g the integrals must exists. The integrals then split by an application of Fubini's theorem, since independence implies that the joint distribution of the variables is the product measure of the marginals.

For the converse we can note that indicator functions on all Borel sets are bounded and measurable and get the result immediately by noting that for any Borel set $P(X \in E) = E(1_E)$ (and similarly for Y) so

$$P(X \in E, Y \in D) = E(1_E(X)1_D(Y)) = E(1_E(X))E(1_D(Y)) = P(X \in E)P(Y \in D)$$

as desired. □

The characterization in Theorem 2.1.4 shows the intimate connections between integrals (expectations) and independence.

In practice when given two real-valued random variables, we could be interested in knowing whether the variables are independent or not. One way to rule out independence is by looking at the covariance of the two variables as the following theorem suggests.

Theorem 2.1.5 (Covariance of independent variables)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and X and Y be real-valued random variables. Assume that $E|X|^2 < \infty$ and $E|Y|^2 < \infty$. If $X \perp\!\!\!\perp Y$ then $\mathrm{Cov}(X, Y) = 0$.

Proof.
The moment assumptions ensure that the integrals in the definition of the covariance exist and then the theorem follows from an argument similar to the one given in Theorem 2.1.4. □

The GCM will rely on a conditional variant of the above result, so let us now develop this theory. Conditional independence is a generalization of independence that is expressed in the language of conditional expectations as defined in the appendix.

Conditional expectations allow us to define conditional probabilities by setting $P(F \mid \mathbb{D}) := E(1_F \mid \mathbb{D})$ for any sub-σ-algebra $\mathbb{D}$ of $\mathbb{F}$ and any $F \in \mathbb{F}$. This is analogous to the usual result that $P(F) = E(1_F)$. We can now define conditional independence.

Definition 2.1.6 (Conditional independence of σ-algebras)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{F}_1$, $\mathbb{F}_2$ and $\mathbb{F}_3$ be sub-σ-algebras of $\mathbb{F}$. If

$$P(F_1 \cap F_2 \mid \mathbb{F}_3) = P(F_1 \mid \mathbb{F}_3) P(F_2 \mid \mathbb{F}_3), \quad \forall F_1 \in \mathbb{F}_1, F_2 \in \mathbb{F}_2,$$

holds almost surely, we say that $\mathbb{F}_1$ *is conditionally independent of* $\mathbb{F}_2$ *given* $\mathbb{F}_3$ and write $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3$.

If $\mathbb{F}_4$ is a fourth sub-σ-algebra of $\mathbb{F}$, we write $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3, \mathbb{F}_4$ as short-hand for $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \sigma(\mathbb{F}_3, \mathbb{F}_4)$.

The definition above looks very similar to the definition of independence (and does in fact contain it by setting $\mathbb{F}_3 = \{\Omega, \varnothing\}$) but there is also an equivalent definition that often becomes helpful.

Theorem 2.1.7 (Equivalent definition of conditional independence)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{F}_1$, $\mathbb{F}_2$ and $\mathbb{F}_3$ be sub-σ-algebras of $\mathbb{F}$. $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3$ if and only if

$$P(F_1 \mid \mathbb{F}_2, \mathbb{F}_3) = P(F_1 \mid \mathbb{F}_3),$$

for all $F_1 \in \mathbb{F}_1$.

Proof.
See the appendix Theorem A.3.1. □

We can also consider conditional independence of random variables.

Definition 2.1.8 (Conditional independence of random variables)**.** Let X and Y be random variables defined on the same probability space $(\Omega, \mathbb{F}, P)$ with values in the measurable spaces $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{G})$ respectively. Let further $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. We say that X *is conditionally independent of* Y *given* $\mathbb{D}$ if $\sigma(X) \perp\!\!\!\perp \sigma(Y) \mid \mathbb{D}$ and write $X \perp\!\!\!\perp Y \mid \mathbb{D}$.

If Z is a third random variable defined on the same probability space with values some measurable space, we define $X \perp\!\!\!\perp Y \mid Z$ to mean $X \perp\!\!\!\perp Y \mid \sigma(Z)$.

If W is a fourth random variable defined on the same probability space with values in some measurable space, we define $X \perp\!\!\!\perp Y \mid Z, W$ to mean $X \perp\!\!\!\perp Y \mid \sigma(Z, W)$.

The interpretation of conditional independence is slightly more subtle than regular independence. If $X \perp\!\!\!\perp Y \mid Z$, then knowing the outcome of X tells me nothing about the outcome of Y if we also know the outcome of Z. This intuition can be seen clearly in the equivalent definition of conditional independence given earlier.

To get an idea of how conditional independence of several variables interacts, we will derive some simple properties of conditional independence.

Theorem 2.1.9 (Fundamental properties of conditional independence)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{F}_1$, $\mathbb{F}_2$, $\mathbb{F}_3$ and $\mathbb{F}_4$ be sub-σ-algebras of $\mathbb{F}$. Then

1. $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3 \iff \mathbb{F}_2 \perp\!\!\!\perp \mathbb{F}_1 \mid \mathbb{F}_3$ (symmetry)
2. $\mathbb{F}_1 \perp\!\!\!\perp (\mathbb{F}_2, \mathbb{F}_3) \mid \mathbb{F}_4 \implies \mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_4 \wedge \mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_3 \mid \mathbb{F}_4$ (decomposition)
3. $\mathbb{F}_1 \perp\!\!\!\perp (\mathbb{F}_2, \mathbb{F}_3) \mid \mathbb{F}_4 \implies \mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid (\mathbb{F}_3, \mathbb{F}_4)$ (weak union)
4. $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3 \wedge \mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_4 \mid (\mathbb{F}_2, \mathbb{F}_3) \implies \mathbb{F}_1 \perp\!\!\!\perp (\mathbb{F}_2, \mathbb{F}_4) \mid \mathbb{F}_3$ (contraction)

Proof.
Symmetry is obvious from the definition of conditional independence and decomposition is also straightforward since both $\mathbb{F}_2$ and $\mathbb{F}_3$ are subsets of $\sigma(\mathbb{F}_2, \mathbb{F}_3)$.

Weak union holds since taking $F_1 \in \mathbb{F}_1$, we get

$$E(1_{F_1} \mid \mathbb{F}_2, \mathbb{F}_3, \mathbb{F}_4) = E(1_{F_1} \mid \mathbb{F}_4),$$

by assumption and since also by decomposition $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_3 \mid \mathbb{F}_4$, we can continue and write

$$E(1_{F_1} \mid \mathbb{F}_4) = E(1_{F_1} \mid \mathbb{F}_3, \mathbb{F}_4),$$

thus proving conditional independence by the equivalent definition given earlier. Finally for contraction, we get

$$E(1_{F_1} \mid \mathbb{F}_2, \mathbb{F}_3, \mathbb{F}_4) = E(1_{F_1} \mid \mathbb{F}_2, \mathbb{F}_3) = E(1_{F_1} \mid \mathbb{F}_3),$$

by $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_4 \mid (\mathbb{F}_2, \mathbb{F}_3)$ and $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3$ respectively, again proving conditional independence by the equivalent definition. □

It is worthwhile to note that conditional independence is preserved under measurable functions.

Theorem 2.1.10 (Conditional independence of functions of random variables)**.** Let X and Y be random variables defined on the same probability space $(\Omega, \mathbb{F}, P)$ with values in the measurable spaces $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{G})$ respectively. Let further $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$ and $f : \mathcal{X} \to \tilde{\mathcal{X}}$ and $g : \mathcal{Y} \to \tilde{\mathcal{Y}}$ be measurable functions into the measurable spaces $(\tilde{\mathcal{X}}, \tilde{\mathbb{E}})$ and $(\tilde{\mathcal{Y}}, \tilde{\mathbb{G}})$ respectively. If $X \perp\!\!\!\perp Y \mid \mathbb{D}$, then $f(X) \perp\!\!\!\perp g(Y) \mid \mathbb{D}$.

Proof.
Note that any $F \in \sigma(f(X))$ is of the form $(f(X) \in E)$ for some $E \in \tilde{\mathbb{E}}$, which is equivalent to $(X \in f^{-1}(E))$. Measurability of f implies that $f^{-1}(E) \in \mathbb{E}$, so the set $(X \in f^{-1}(E))$ is in $\sigma(X)$. A similar argument can be performed on sets in $\sigma(g(Y))$. This proves the result since we know that sets in $\sigma(X)$ and $\sigma(Y)$ satisfy the criterion required for conditional independence. □

Just as it was done for independence, we can characterize conditional independence.

Theorem 2.1.11 (Characterization of conditional independence)**.** Let X and Y be random variables defined on the same probability space $(\Omega, \mathbb{F}, P)$ with values in the measurable spaces $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{G})$ respectively. Let further $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. Then for all $f : \mathcal{X} \to \mathbb{R}$, $g : \mathcal{Y} \to \mathbb{R}$ Borel measurable and bounded functions we have

$$E(f(X)g(Y) \mid \mathbb{D}) = E(f(X) \mid \mathbb{D})E(g(Y) \mid \mathbb{D})$$

if and only if $X \perp\!\!\!\perp Y \mid \mathbb{D}$.

Proof.
See [5] Proposition 2.3.28. □

We can use the characterization in Theorem 2.1.11 to show that the conditional expectation of a product of integrable real-valued variables factorizes when they are conditionally independent.

Theorem 2.1.12 (Conditional expectation of conditionally independent variables factorizes)**.** Let X and Y be real-valued random variables defined on the same probability space $(\Omega, \mathbb{F}, P)$. Let further $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. Assume that X, Y and XY are integrable. Then if $X \perp\!\!\!\perp Y \mid \mathbb{D}$,

$$E(XY \mid \mathbb{D}) = E(X \mid \mathbb{D})E(Y \mid \mathbb{D}).$$

Proof.
Note that we can define a sequence of bounded and measurable functions $f_n(x) = 1_{(|x| \leqslant n)}(x)$ such that f_n converges to the identity. It is straightforward to see that $f_n(xy) = f_n(x)f_n(y)$. Then

$$E(XY \mid \mathbb{D}) = E\left(\lim_{n\to\infty} f_n(XY) \mid \mathbb{D}\right) = \lim_{n\to\infty} E(f_n(XY) \mid \mathbb{D}),$$

by the conditional dominated convergence theorem since $f_n(XY)$ is bounded by XY which is integrable by assumption. Continuing we get by conditional independence and the boundedness of f_n

$$\lim_{n\to\infty} E(f_n(XY) \mid \mathbb{D}) = \lim_{n\to\infty} E(f_n(X)f_n(Y) \mid \mathbb{D}) = \lim_{n\to\infty} E(f_n(X) \mid \mathbb{D})E(f_n(Y) \mid \mathbb{D}).$$

By again applying the conditional dominated convergence theorem (see Theorem A.1.26) since $f_n(X)$ is bounded by integrable X and similarly for Y, we get the desired result. □

In practice when we are given three real-valued random variables X, Y and Z, we would like to find a way to determine whether $X \perp\!\!\!\perp Y \mid Z$. To that end we can define a conditional variant of covariance.

Definition 2.1.13 (Conditional covariance)**.** Let X and Y be real-valued random variables defined on a probability space $(\Omega, \mathbb{F}, P)$ and let $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. Assume that $E|X|^2 < \infty$ and $E|Y|^2 < \infty$. We define the *conditional covariance of X and Y given $\mathbb{D}$* as

$$\mathrm{Cov}(X, Y \mid \mathbb{D}) = E\left([X - E(X \mid \mathbb{D})][Y - E(Y \mid \mathbb{D})] \mid \mathbb{D}\right).$$

Applying simple laws for conditional expectations reveals that

$$\mathrm{Cov}(X, Y \mid \mathbb{D}) = E(XY \mid \mathbb{D}) - E(X \mid \mathbb{D})E(Y \mid \mathbb{D}).$$

Note that the conditional covariance is a random variable and not simply a real number. Just as the covariance of independent random variables is zero the conditional covariance of conditionally independent random variables is also zero.

Theorem 2.1.14 (Conditional covariance of conditionally independent variables)**.** Let X and Y be real-valued random variables defined on a probability space $(\Omega, \mathbb{F}, P)$ and let $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. Assume that $E|X|^2 < \infty$ and $E|Y|^2 < \infty$. Then if $X \perp\!\!\!\perp Y \mid \mathbb{D}$ we have $\mathrm{Cov}(X, Y \mid \mathbb{D}) = 0$.

Proof.
Follows immediately from 2.1.12. □

The following theorem forms the basis of the GCM.

Theorem 2.1.15 (Product of residuals of conditionally independent variables is zero)**.** Let X and Y be real-valued random variables defined on a probability space $(\Omega, \mathbb{F}, P)$ and let $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. Assume that $E|X|^2 < \infty$ and $E|Y|^2 < \infty$.

Define the residuals $\varepsilon = X - E(X \mid \mathbb{D})$ and $\xi = Y - E(Y \mid \mathbb{D})$.

Then if $X \perp\!\!\!\perp Y \mid \mathbb{D}$ we have $E(\xi\varepsilon) = 0$.

Proof.
Note that by the tower property it is sufficient to show that $E(\varepsilon\xi \mid \mathbb{D}) = 0$. This is know immediate from the definition of the conditional cross-covariance and Theorem 2.1.14, since

$$E(\varepsilon\xi \mid \mathbb{D}) = E([X - E(X \mid \mathbb{D})][Y - E(Y \mid \mathbb{D})] \mid \mathbb{D}) = \mathrm{Cov}(X, Y \mid \mathbb{D}).$$

□

When given n i.i.d. observations of three random variables X, Y and Z, the GCM will be based on estimating the conditional expectations of X given Z and Y given Z, forming the residuals and testing whether the mean of the product of the residuals is zero. By Theorem 2.1.15 if that is not the case, X and Y are not conditionally independent given Z. We will expand on this later.

We can in fact generalize Theorem 2.1.15 to give yet another characterization of conditional independence.

Theorem 2.1.16 (Daudin's lemma)**.** Let X, Y and Z be real-valued random variables defined on a probability space $(\Omega, \mathbb{F}, P)$. Then $X \perp\!\!\!\perp Y \mid Z$ if and only if

$$E(f(X, Z)g(Y, Z)) = 0$$

for all f, g measurable and real-valued with $E(f(X, Z) \mid Z) = E(g(Y, Z) \mid Z) = 0$, $E(f(X, Z)^2) < \infty$ and $E(g(Y, Z)^2) < \infty$.

Proof.
A proof is given in [7], where the conditional independence statement $X \perp\!\!\!\perp Y \mid Z$ is defined as $E(f(X, Z)g(Y, Z) \mid Z) = E(f(X, Z) \mid Z)E(g(Y, Z) \mid Z)$. Using the definitions of this thesis, this would be $(X, Z) \perp\!\!\!\perp (Y, Z) \mid Z$ but we show in Theorem A.3.2 that this is equivalent to assuming $X \perp\!\!\!\perp Y \mid Z$. □

Daudin's lemma is a strengthening of Theorem 2.1.15, since ε and ξ are examples of square integrable functions with conditional mean zero. Using Theorem 2.1.15 to construct a test, we would not be able to detect all cases of conditional dependence but Daudin's lemma states that if we performed the test on all possible transformations of the random variables, we would be able to detect all cases of conditional dependence. This is a theoretical consideration, since of course such a test would not be possible in practice.

The standard modes of convergence of sequences of real-valued random variables are revised in the appendix. We will also apply some uniform variants of convergence for families of sequences of random variables.

Definition 2.1.17 (Uniform convergence of random variables)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let Θ be some parameter space (think of Θ being a subset of $\mathbb{R}^d$ or even a set of probability measures). Let further $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be a family of real-valued random variables defined on $(\Omega, \mathbb{F}, P)$. Let $(X_\theta)_{\theta\in\Theta}$ be another family of real-valued random variables. Then

1. If for every $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} P(|X_{n,\theta} - X_\theta| \geqslant \varepsilon) = 0,$$

   we say that $X_{n,\theta}$ *converges to* X_θ *in probability uniformly in* θ and write $X_{n,\theta} \overset{P}{\rightrightarrows}_\Theta X_\theta$.

2. If for every bounded, continuous, real-valued function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} |E(f(X_{n,\theta})) - E(f(X_\theta))| = 0,$$

   we say that $X_{n,\theta}$ *converges to* X_θ *in distribution uniformly in* θ and write $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows}_\Theta X_\theta$.

We will sometimes omit the subscripted Θ from the notation when it is clear from the context. We will also abuse notation slightly and write $X_{n,\theta} \rightrightarrows X$ where X is a single random variable, by which we mean that $X_{n,\theta}$ converges uniformly to the family $X_\theta = X$ for all $\theta \in \Theta$.

Each family $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ is often thought of as a sequence for each θ, i.e. for each $\theta_0 \in \Theta$, we would consider $(X_{n,\theta_0})_{n\in\mathbb{N}}$ when usually thinking about convergence of random variables. The uniform definitions allow us to consider what happens to convergence across multiple possible distributions of a sequence simultaneously.

It is quite easy to see that convergence in distribution uniformly in θ implies convergence in distribution for every θ and similarly for convergence in probability. These forms of convergence turn out to be a natural language to phrase various requirements on tests to ensure that they are uniformly well-behaved across all possible distributions.

Uniform convergence is preserved under continuous transformations as seen in the following theorem.

Theorem 2.1.18 (Uniform continuous mapping theorem)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be a family of real-valued random variables defined on $(\Omega, \mathbb{F}, P)$. Let $(X_\theta)_{\theta\in\Theta}$ be another family of real-valued variables on the same space and assume that $X_{n,\theta} \overset{\mathcal{D}}{\to}_\Theta X_\theta$. Let $h : \mathbb{R} \to \mathbb{R}$ be a continuous mapping. Then

$$h(X_{n,\theta}) \overset{\mathcal{D}}{\to}_\Theta h(X_\theta).$$

Proof.
Note that if f is a continuous, bounded and real-valued function, then so is $f \circ h$, thus the result follows immediately. □

Many of the usual properties of convergence in distribution boil down to the question of whether a given sequence or family is tight.

Definition 2.1.19 (Tightness of a family of random variables)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $X_{a\in A}$ be a family of real-valued random variables on $(\Omega, \mathbb{F}, P)$ indexed by the set A. We say that $X_{a\in A}$ is *tight* if for all $\varepsilon > 0$, there exists some $M > 0$ so that

$$\sup_{a\in A} P(|X_a| > M) < \varepsilon.$$

Both a single measure and any sequence of random variables converging in distribution are tight, which is applied when proving theorems such as Slutsky's theorem. Neither a single family $(X_\theta)_{\theta\in\Theta}$ nor a family $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ converging uniformly in distribution are a priori tight. Finding assumptions guaranteeing tightness is non-trivial but assuming that that for each $n \in \mathbb{N}$ the family $(X_{n,\theta})_{\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ are tight is sufficient to prove a version of Slutsky's theorem.

Lemma 2.1.20 (Tightness of uniformly convergent sequence)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be a family of real-valued random variables defined on $(\Omega, \mathbb{F}, P)$. Let $(X_\theta)_{\theta\in\Theta}$ be another family of real-valued variables on the same space and assume that for each $n \in \mathbb{N}$ the family $(X_{n,\theta})_{\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ are tight.

Then if for every bounded, uniformly continuous, real-valued function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} |E(f(X_{n,\theta})) - E(f(X_\theta))| = 0,$$

the family $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ is tight.

Proof.
See [27] Lemma 3.1.6 for a proof in the non-uniform case and note that the extra assumptions of tightness of $(X_{n,\theta})_{\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ allow for the proof to also hold in the uniform case. □

This lets us conclude that it is sufficient to consider uniformly continuous test functions when proving uniform convergence in distribution.

Theorem 2.1.21 (Uniform convergence in distribution and uniform continuity)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be a family of real-valued random variables defined on $(\Omega, \mathbb{F}, P)$. Let $(X_\theta)_{\theta\in\Theta}$ be another family of real-valued variables on the same space and assume that for each $n \in \mathbb{N}$ the family $(X_{n,\theta})_{\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ are tight. Then $X_{n,\theta} \overset{\mathcal{D}}{\to}_\Theta X_\theta$ if and only if for every bounded, uniformly continuous, real-valued function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} |E(f(X_{n,\theta})) - E(f(X_\theta))| = 0.$$

Proof.
See [27] Theorem 3.1.7 for a proof in the non-uniform case and the proof then follows in the uniform case applying Lemma 2.1.20 instead of the non-uniform version in the proof. □

We will apply Theorem 2.1.21 to prove Slutsky's lemma.

Theorem 2.1.22 (Slutsky's lemma for uniform convergence)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ and $(Y_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be two families of real-valued random variables defined on $(\Omega, \mathbb{F}, P)$. Let $(X_\theta)_{\theta\in\Theta}$ be another family of real-valued variables on the same space and assume that for each $n \in \mathbb{N}$ the family $(X_{n,\theta})_{\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ are tight. Assume further that $X_{n,\theta} \overset{\mathcal{D}}{\to} X_\theta$ and $Y_{n,\theta} \overset{P}{\to} 0$. Then

$$X_{n,\theta} + Y_{n,\theta} \overset{\mathcal{D}}{\to} X_\theta.$$

Proof.
By Theorem 2.1.21 it suffices to prove

$$\sup_{\theta\in\Theta} |E(f(X_{n,\theta} + Y_{n,\theta})) - E(f(X_\theta))| \to 0$$

as $n \to \infty$ for all bounded, uniformly continuous real-valued functions f.

Note that

$$\begin{aligned}
&\sup_{\theta\in\Theta} |E(f(X_{n,\theta} + Y_{n,\theta})) - E(f(X_\theta))| \\
&\leqslant \sup_{\theta\in\Theta} |E(f(X_{n,\theta} + Y_{n,\theta})) - E(f(X_{n,\theta}))| + \sup_{\theta\in\Theta} |E(f(X_{n,\theta})) - E(f(X_\theta))|
\end{aligned}$$

and the second term goes to 0 by assumption, so it suffices to show that the first term goes to 0. By the triangle inequality for integrals and linearity of the integral we get

$$\sup_{\theta\in\Theta}|E(f(X_{n,\theta}+Y_{n,\theta}))-E(f(X_{n,\theta}))|\leqslant\sup_{\theta\in\Theta}E|f(X_{n,\theta}+Y_{n,\theta})-f(X_{n,\theta})|.$$

For any $\varepsilon>0$, we can find $\delta>0$ from the uniform continuity of f so that $|f(x+y)-f(x)|<\varepsilon$ for all $|y|<\delta$. This lets us partition the integral into a region where $|Y_{n,\theta}|\leqslant\delta$ (and thus where $|f(X_{n,\theta}+Y_{n,\theta})-f(X_{n,\theta})|\leqslant\varepsilon$) and a region where $|Y_{n,\theta}|>\delta$. We get by also applying the triangle inequality to the second integral that

$$\begin{aligned}\sup_{\theta\in\Theta}E|f(X_{n,\theta}+Y_{n,\theta})-f(X_{n,\theta})|&\leqslant\varepsilon+\sup_{\theta\in\Theta}E[1_{(|Y_{n,\theta}|>\delta)}(|f(X_{n,\theta}+Y_{n,\theta})|+|f(X_{n,\theta}|)]\\&\leqslant\varepsilon+\sup_{\theta\in\Theta}2\|f\|_\infty P(|Y_{n,\theta}|>\delta),\end{aligned}$$

where $\|f\|_\infty=\sup_{x\in\mathbb{R}}f(x)$, which is finite by assumption, so the second term can be made arbitrarily small by choosing a large enough n. Since ε was arbitrary, we are done. □

Unfortunately we are not able to continue this development and generalize to the usual Slutsky's theorem under the assumptions given here. Bengs and Holzmann use stronger assumptions in [1] to get the following result.

Theorem 2.1.23 (Slutsky's theorem for uniform convergence)**.** Let $(\Omega,\mathbb{F},P)$ be a probability space, let Θ be some parameter space and let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ and $(Y_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be two families of real-valued random variables defined on $(\Omega,\mathbb{F},P)$. Let $(X_\theta)_{\theta\in\Theta}$ be another family of real-valued variables on the same space, let $(y_\theta)_{\theta\in\Theta}$ be a family of real numbers and assume that $X_{n,\theta}\overset{\mathcal{D}}{\to}X_\theta$ and $Y_{n,\theta}\overset{P}{\to}y_\theta$.

Assume further that the family of measures $(X_\theta(P))_{\theta\in\Theta}$ is uniformly absolutely continuous with respect to some continuous probability measure Q on $(\mathbb{R},\mathbb{B})$, i.e. for any $\varepsilon>0$, there exists $\delta>0$, such that for any $B\in\mathbb{B}$ with $Q(B)<\delta$ we have

$$\sup_{\theta\in\Theta}P(X_\theta\in B)<\varepsilon.$$

Then

$$X_{n,\theta}+Y_{n,\theta}\overset{\mathcal{D}}{\to}X_\theta+y_\theta$$

and

$$X_{n,\theta}Y_{n,\theta}\overset{\mathcal{D}}{\to}X_\theta y_\theta.$$

Proof.
See [1] Theorem 6.3. □

We will solely apply this full version of Slutsky's theorem in the case where X_θ is a family of normal distributions with mean zero and bounded variances and we now prove that such a family is in fact uniformly absolutely continuous with respect to a continuous measure on $\mathbb{R}$.

Theorem 2.1.24 (Mean zero normal distributions are uniformly absolutely continuous)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let $(X_\theta)_{\theta\in\Theta}$ be a family of real-valued random variables. Let $\sigma(\theta)$ be a function such that $\sigma(\theta)$ is bounded and bounded away from zero for all θ and assume that $X_\theta \sim \mathcal{N}(0, \sigma(\theta)^2)$. Then we assume that $(X_\theta)_{\theta\in\Theta}$ is uniformly absolutely continuous with respect to some continuous probability measure on $(\mathbb{R}, \mathbb{B})$.

Proof.
Let $\sigma^2_{\sup} := \sup_{\theta\in\Theta} \sigma^2(\theta)$ and $\sigma^2_{\inf}$ similarly. We intend to show that the family $(P_\theta)_{\theta\in\Theta} = (X_\theta(P))_{\theta\in\Theta}$ is uniformly absolutely continuous with respect to the measure $Q = \mathcal{N}(0, \sigma^2_{\sup})$. To that end let $\varepsilon > 0$ be given and choose first $M > 0$ from the tightness of Q so that

$$Q([-M, M]^c) < \frac{\varepsilon}{2}.$$

Note first that $P_\theta([-M, M]^c) \leqslant Q([-M, M]^c)$ for any θ. We can see this by arguing that $P_\theta([-M, M]) \geqslant Q([-M, M])$, which can be seen by performing integration by substitution

$$P_\theta([-M, M]) = \int_{-M}^{M} \frac{1}{\sqrt{2\pi\sigma(\theta)^2}} e^{\frac{-x^2}{2\sigma^2(\theta)}}\,\mathrm{d}x = \int_{-\frac{\sigma^2_{\sup}}{\sigma^2(\theta)}M}^{\frac{\sigma^2_{\sup}}{\sigma^2(\theta)}M} \frac{1}{\sqrt{2\pi\sigma^2_{\sup}}} e^{\frac{-u^2}{2\sigma^2_{\sup}}}\,\mathrm{d}u \geqslant Q([-M, M]),$$

where the final equality is due to $\frac{\sigma^2_{\sup}}{\sigma^2(\theta)} \geqslant 1$. Note also that if $Q(A) < \delta$ for some $\delta > 0$ then $m(A \cap [-M, M]) < \frac{\delta}{\varphi_Q(M)}$ where m denotes the Lebesgue measure and φ_Q is the density of Q. This has to holds since if not then

$$Q(A) \geqslant Q(A \cap [-M, M]) = \int_{A\cap[-M,M]} \varphi_Q(x)\,\mathrm{d}x \geqslant \varphi_Q(M) m(A \cap [-M, M]) \geqslant \delta$$

since $\varphi_Q(M)$ is the smallest value that φ_Q attains over $[-M, M]$. Let $C = \sup_{\theta\in\Theta} \varphi_\theta(0)$ where φ_θ is the density of P_θ with respect to the Lebesgue measure and set $\delta = \min\left(\frac{\varepsilon}{2}, \frac{\varepsilon}{2}\frac{\varphi_Q(M)}{C}\right)$. Then for any $\theta \in \Theta$ and $A \in \mathbb{B}$ with $Q(A) < \delta$, we have

$$\begin{aligned} P_\theta(A) &= P_\theta(A \cap [-M, M]^c) + P_\theta(A \cap [-M, M]) \leqslant Q([-M, M]^c) + Cm(A \cap [-M, M]) \\ &\leqslant \frac{\varepsilon}{2} + \frac{\delta}{\varphi_Q(M)} C \leqslant \varepsilon \end{aligned}$$

finishing the proof.

□

We have a version of the Law of Large numbers in the context of uniform convergence.

Theorem 2.1.25 (Uniform Law of Large Numbers)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let $(X_\theta)_{\theta\in\Theta}$ be a real-valued family of random variables defined on $(\Omega, \mathbb{F}, P)$. Assume that there exists $\eta > 0$ so that $\sup_{\theta\in\Theta} E|X_\theta|^{1+\eta} < \infty$ and let $\mu(\theta) = E(X_\theta)$. Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be a family of real-valued random variables such that for each $\theta_0 \in \Theta$ the sequence $(X_{n,\theta_0})_{n\in\mathbb{N}}$ is independent and with the same distribution as X_{θ_0}. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta} \overset{P}{\underset{\Theta}{\to}} \mu(\theta).$$

Proof.
Assume without loss of generality that $\mu(\theta) = 0$, since if the result holds, we can instead consider $\tilde{X}_{n,\theta} = X_{n,\theta} - \mu(\theta)$.

Let $\varepsilon > 0$ be given and note that for every $M > 0$, we can write

$$\begin{aligned}
&\sup_{\theta\in\Theta} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta}\right| \geqslant \varepsilon\right) \\
&\leqslant \sup_{\theta\in\Theta} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta}1_{(|X_{i,\theta}|\leqslant M)}\right| \geqslant \varepsilon\right) + \sup_{\theta\in\Theta} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta}1_{(|X_{i,\theta}|> M)}\right| \geqslant \varepsilon\right).
\end{aligned}$$

For the first term, note that by $1+\eta$-order Markov's inequality, the triangle inequality and the i.i.d nature of $X_{n,\theta}$ for each θ yields

$$\sup_{\theta\in\Theta} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta}1_{(|X_{i,\theta}|\leqslant M)}\right| \geqslant \varepsilon\right) \leqslant \sup_{\theta\in\Theta} \frac{E(X_\theta^{1+\eta}1_{(|X_\theta|\leqslant M)})}{n^\eta\varepsilon^{1+\eta}} \leqslant \frac{M^{1+\eta}}{n^\eta\varepsilon^{1+\eta}}.$$

For the second term, by the the arguments as above but using first order Markov's inequality, we get

$$\sup_{\theta\in\Theta} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta}1_{(|X_{i,\theta}|> M)}\right| \geqslant \varepsilon\right) \leqslant \sup_{\theta\in\Theta} \frac{E(|X_\theta|1_{(|X_\theta|> M)})}{\varepsilon}.$$

Using Hölder's inequality on the integral, we get

$$E(|X_\theta|1_{(|X_\theta|>M)} \leqslant E(|X_\theta|^{1+\eta})P(|X_\theta| > M)^{\frac{1+\eta}{\eta}}.$$

We can bound the probability by Markov's inequality once again and get

$$P(|X_\theta| > M) \leqslant \frac{E|X_\theta|}{M},$$

which implies that

$$\sup_{\theta\in\Theta} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta}1_{(|X_{i,\theta}|>M)}\right| \geqslant \varepsilon\right) \leqslant \frac{1}{\varepsilon M^{(1+\eta)/\eta}} \sup_{\theta} E(|X_\theta|^{1+\eta})E(|X_\theta|)^{(1+\eta)/\eta}.$$

We can thus choose M sufficiently large to ensure that the second term is small and then choose n sufficiently large, so that the first term becomes small. □

We also have a central limit theorem.

Theorem 2.1.26 (Uniform Central Limit Theorem)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let $(X_\theta)_{\theta\in\Theta}$ be a real-valued family of random variables defined on $(\Omega, \mathbb{F}, P)$. Assume that there exists $\eta > 0$ so that $\sup_{\theta\in\Theta} E|X_\theta|^{2+\eta} < \infty$ and let $\mu(\theta) = E(X_\theta)$ and $\sigma^2(\theta) = \mathrm{Var}(X_\theta)$. Assume further that $\inf_\theta \sigma^2(\theta) > 0$. Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be a family of real-valued random variables such that for each $\theta_0 \in \Theta$ the sequence $(X_{n,\theta_0})_{n\in\mathbb{N}}$ is independent and with the same distribution as X_{θ_0}. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_{i,\theta} - \mu(\theta)) \overset{\mathcal{D}}{\to}_\Theta \mathcal{N}(0, \sigma^2(\theta)).$$

Proof.
We can assume that $\mu(\theta) = 0$, since otherwise we consider the variables $\tilde{X}_{n,\theta} = X_{n,\theta} - \mu(\theta)$. Define

$$W_{n,\theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_{i,\theta} - \mu(\theta))$$

and $Z_\theta \sim \mathcal{N}(0, \sigma^2(\theta)$.

Note that to show uniform convergence in distribution, it is sufficient to show that for every sequence $(\theta_n)_{n\in\mathbb{N}}$ in Θ, we have

$$|E(f(W_{n,\theta_n})) - E(f(Z_{\theta_n}))| \to 0$$

as $n \to \infty$. This holds since if

$$\sup_{\theta\in\Theta} |E(f(W_{n,\theta})) - E(f(Z_\theta))| \nrightarrow 0$$

as $n \to \infty$, there would exist $\varepsilon > 0$ and sequences $(\theta_k)_{k\in\mathbb{N}}$ and $(n_k)_{k\in\mathbb{N}}$ so that

$$|E(f(W_{n_k,\theta_k})) - E(f(Z_{\theta_k}))| \geqslant \varepsilon$$

for all $k \in \mathbb{N}$.

Take any sequence $(\theta_n)_{n\in\mathbb{N}}$ in Θ and define

$$Y_{nk} = \frac{1}{\sqrt{n}} \frac{X_{k,\theta_k}}{\sigma(\theta_k)}$$

for $n \in \mathbb{N}$ and $1 \leqslant k \leqslant n$ and

$$S_n = \sum_{k=1}^{n} Y_{nk} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_{i,\theta_i}}{\sigma(\theta_i)}.$$

$(Y_nk)_{1\leqslant k\leqslant n}$ satisfies the conditions of Lyapounov's CLT (see Theorem A.1.20). The rows are independent and centered and the variance of S_n is 1 by construction. The Lyapounov condition is satisfied since

$$\begin{aligned}\lim_{n\to\infty} \sum_{k=1}^{n} E|Y_{nk}|^{2+\eta} &= \lim_{n\to\infty} \frac{1}{\sqrt{n}^{\eta}} \frac{1}{n} \sum_{k=1}^{n} \frac{E|X_{k,\theta_k}|^{2+\eta}}{\sigma^{2+\eta}(\theta_k)} \\ &\leqslant \lim_{n\to\infty} \frac{1}{\sqrt{n}^{\eta}} \frac{1}{\inf_{\theta\in\Theta} \sigma^{2+\eta}(\theta)} \sup_{\theta\in\Theta} E|X_\theta|^{2+\eta} \to 0\end{aligned}$$

as $n \to \infty$.

The above shows that convergence holds for all sequence i.e. that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_{i,\theta}}{\sigma(\theta)} \overset{\mathcal{D}}{\rightrightarrows} \mathcal{N}(0,1).$$

Applying Theorem 2.1.23 with Assumption 2.1.24 will give the desired result, which we can do since the moment conditions given imply tightness by Markov's inequality. □

The approach for uniform convergence results given here is not unique and other approaches using different assumptions to account for tightness can be seen in [1] as mentioned ealier or in [14].

In addition to the properties of uniform convergence derived above, when proving the asymptotic properties of the GCM, we will need the following lemmas.

Lemma 2.1.27. Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let further $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be a family of real-valued random variables defined on $(\Omega, \mathbb{F}, P)$. If $X_{n,\theta} \overset{P}{\rightrightarrows} 0$ and there exists $C > 0$ such that $|X_{n,\theta}| < C$ for all $n \in \mathbb{N}$ and $\theta \in \Theta$, then

$$\sup_{\theta\in\Theta} E(|X_{n,\theta}|) \to 0$$

as $n \to \infty$.

Proof.
Let $\varepsilon > 0$ be given. Note that

$$\begin{aligned}\sup_{\theta\in\Theta} E(|X_{n,\theta}|) &\leqslant \sup_{\theta\in\Theta} E\left(|X_{n,\theta}|1_{(|X_{n,\theta}|<\varepsilon/2)}\right) + \sup_{\theta\in\Theta} E\left(|X_{n,\theta}|1_{(|X_{n,\theta}|\geqslant\varepsilon/2)}\right) \\ &\leqslant \frac{\varepsilon}{2} + C\sup_{\theta\in\Theta} P(|X_{n,\theta}| \geqslant \varepsilon/2).\end{aligned}$$

By assumption for any $\eta > 0$, we can choose $N \in \mathbb{N}$ so that for all $n \geqslant N$, we can make $\sup_{\theta\in\Theta} P(|X_{n,\theta}| \geqslant \varepsilon/2) < \eta$. Thus choosing N to parry $\eta = \frac{\varepsilon}{2C}$, we get

$$\sup_{\theta\in\Theta} E(|X_{n,\theta}|) < \varepsilon.$$

Since ε was arbitrary, we get the desired result. □

Lemma 2.1.28. Let $(\Omega, \mathbb{F}, P)$ be a probability space, let Θ be some parameter space and let further $(X_{n,\theta})_{n\in\mathbb{N}}$ be a family of real-valued random variables defined on $(\Omega, \mathbb{F}, P)$. Let $(X_\theta)_{\theta\in\Theta}$ be another family of real-valued random variables and let $(\mathbb{F}_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be a family of sub-σ-algebras of $\mathbb{F}$. If $E(|X_{n,\theta}| \mid \mathbb{F}_{n,\theta}) \overset{P}{\to}_\Theta 0$ then $X_{n,\theta} \overset{P}{\to}_\Theta 0$.

Proof.
Let $\varepsilon > 0$ be given and note that by Markovs inequality

$$\sup_{\theta\in\Theta} P(|X_{n,\theta}| \geqslant \varepsilon) \leqslant \sup_{\theta\in\Theta} P(|X_{n,\theta}| \wedge \varepsilon \geqslant \varepsilon) \leqslant \sup_{\theta\in\Theta} \frac{E(|X_{n,\theta}| \wedge \varepsilon)}{\varepsilon}.$$

We will be done, if we can show that $\sup_{\theta\in\Theta} E(|X_{n,\theta}| \wedge \varepsilon) \to 0$ as $n \to \infty$. Note that by monotonicity of conditional expectations, for each $\theta \in \Theta$ we have

$$|X_{n,\theta}| \wedge \varepsilon \leqslant \varepsilon \Longrightarrow E(|X_{n,\theta}| \wedge \varepsilon \mid \mathbb{F}_{n,\theta}) \leqslant E(\varepsilon \mid \mathbb{F}_{n,\theta}) = \varepsilon,$$

and

$$|X_{n,\theta}| \wedge \varepsilon \leqslant |X_{n,\theta}| \Longrightarrow E(|X_{n,\theta}| \wedge \varepsilon \mid \mathbb{F}_{n,\theta}) \leqslant E(X_{n,\theta} \mid \mathbb{F}_{n,\theta}).$$

Combining both of the above expressions, we get

$$E(|X_{n,\theta}| \wedge \varepsilon \mid \mathbb{F}_{n,\theta}) \leqslant E(|X_{n,\theta}| \wedge \varepsilon \mid \mathbb{F}_{n,\theta}) \wedge \varepsilon.$$

This lets us write by the tower property and monotonicity of integrals

$$\sup_{\theta\in\Theta} E(|X_{n,\theta}| \wedge \varepsilon) = \sup_{\theta\in\Theta} E(E(|X_{n,\theta}| \wedge \varepsilon \mid \mathbb{F}_{n,\theta})) \leqslant \sup_{\theta\in\Theta} E(E(|X_{n,\theta}| \mid \mathbb{F}_{n,\theta}) \wedge \varepsilon).$$

Now the conclusion follows from the assumptions and Lemma 2.1.27 since $E(|X_{n,\theta}| \mid \mathbb{F}_{n,\theta}) \wedge \varepsilon$ goes to 0 in probability uniformly in $\theta \in \Theta$ and is bounded by ε. □

2.2 Statistical models and test theory

In this section, we present the fundamentals of statistical models and test theory. We will mainly follow the development in [8] and [13].

In statistics we observe the outcome of some random variable X with values in a set $\mathcal{X}$, which has unknown distribution $X(P)$ ($X(P)$ is the push-forward measure of P under X). A statistical model consists of a set of possible distributions for X and we attempt to determine which of these distributions are acceptable (or rather which are unacceptable) for the random phenomenon that is being modelled.

Definition 2.2.1 (Statistical model)**.** A *statistical model* consists of a *sample space*, $\mathcal{X}$, a σ-algebra defined on $\mathcal{X}$, $\mathbb{E}$, and a set of probability measures on $(\mathcal{X}, \mathbb{E})$, $\mathcal{P}$.

Remark 2.2.2 (Observations and sampling assumptions)**.** Throughout this thesis we will always work under the assumption of i.i.d. sampling. We will assume that we observe a sequence $(x_i)_{i\in\mathbb{N}}$ of observations from a sequence of independent and identically distributed copies of X, $(X_i)_{i\in\mathbb{N}}$. We let $X^{(n)}$ denote the joint distribution of the first n of these copies.

A simple example of a statistical model is $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$, $\mathbb{E}$ as the Borel σ-algebra on $\mathbb{R}^d$ and $\mathcal{P}$ as the set of normal distributions on $\mathbb{R}^d$ with unknown mean and covariance.

To draw inference i.e. deciding if there are some $\nu \in \mathcal{P}$ that match observation more than others, statisticians work with the concept of a hypothesis.

Definition 2.2.3 (Hypothesis)**.** Let $(\mathcal{X}, \mathbb{E}, \mathcal{P})$ be a statistical model. A *hypothesis*, H_0, is a subset $\mathcal{P}_0$ of the full family of probability measures, $\mathcal{P}$. The *alternative hypothesis*, H_1, to H_0 is the complement $\mathcal{P}_1 = \mathcal{P}\backslash\mathcal{P}_0$. A hypothesis is called *simple* if it consists of a single measure and *composite* otherwise.

The interpretation of a hypothesis is that the true data-generating mechanism is in the set $\mathcal{P}_0$. Most of classical statistics is built upon the idea of observing an outcome, choosing a suitable model and constructing hypotheses within the model to predict and understand the phenomenon. To formalize the process of accepting and rejecting hypotheses, statisticians work with concept of a test.

Definition 2.2.4 (Test of hypothesis)**.** Let $(\mathcal{X}, \mathbb{E}, \mathcal{P})$ be a statistical model and H_0 a hypothesis. We define a *test* as a sequence of partitions of $\mathcal{X}^n$ into an *acceptance region* $\mathcal{A}_n$ and a *critical region* $\mathcal{A}_n^c$. This partition is also expressed through the sequence of *critical functions* $\psi_n : \mathcal{X}^n \to \{0, 1\}$ given by

$$\psi_n(x) = \begin{cases} 0 & \text{if } x \in \mathcal{A}_n \\ 1 & \text{if } x \in \mathcal{A}_n^c \end{cases}.$$

A test is either $(\mathcal{A}_n)_{n\in\mathbb{N}}$ or equivalently $(\psi_n)_{n\in\mathbb{N}}$.

We will refer to both $\mathcal{A}_n$ and ψ_n as tests, however we will primarily use the functional definition. Given a sample of size n from a statistical model and when considering a hypothesis, we simply apply the n'th critical function to the observation and if $\psi_n(x) = 0$, we accept the hypothesis and if $\psi_n(x) = 1$ we reject it.

The definition above is abstract and allows for poor tests (we could for instance reject or accept everything and this would still be valid tests). We now define some properties of tests that helps us determine whether they are useful in testing a hypothesis.

Definition 2.2.5 (Properties of tests)**.** Let $(\psi_n)_{n\in\mathbb{N}}$ be a sequence of tests of a hypothesis with null set of probability measures $\mathcal{P}_0$. For a given level $\alpha \in (0,1)$, we say that

1. the sequence $(\psi_n)_{n\in\mathbb{N}}$ has *valid level* if for every n
$$\sup_{\nu\in\mathcal{P}_0} P_\nu(\psi_n = 1) \leqslant \alpha,$$
2. the sequence $(\psi_n)_{n\in\mathbb{N}}$ has *uniformly asymptotic level* if
$$\limsup_{n\to\infty} \sup_{\nu\in\mathcal{P}_0} P_\nu(\psi_n = 1) \leqslant \alpha,$$
3. the sequence $(\psi_n)_{n\in\mathbb{N}}$ has *pointwise asymptotic level* if
$$\sup_{\nu\in\mathcal{P}_0} \limsup_{n\to\infty} P_\nu(\psi_n = 1) \leqslant \alpha,$$

where P_ν is short-hand for the probability assuming that $X(P) \sim \nu$.

A test holding level is a way of ensuring, that we do not reject true hypotheses too often. Some of these properties imply each other as the following theorem shows.

Proposition 2.2.6 (Relations between properties of tests)**.** Let $(\mathcal{X}, \mathbb{E}, \mathcal{P})$ be a statistical model, let H_0 be a hypothesis and let $(\psi_n)_{n\in\mathbb{N}}$ be a sequence of tests for the hypothesis.

1. If the sequence $(\psi_n)_{n\in\mathbb{N}}$ has valid level, then it also has uniformly asymptotic level.
2. If the sequence $(\psi_n)_{n\in\mathbb{N}}$ has uniformly asymptotic level then it also has pointwise asymptotic level.

Proof.

1. Trivial, since lim sup respects inequalities.

2. Let

$$m_n = \sup_{\nu \in \mathcal{P}_0} P_\nu(\psi_n = 1).$$

Then clearly $P_\nu(\psi_n = 1) \leqslant m_n$ for every n and ν. Thus since lim sup respects inequalities

$$\limsup_{n\to\infty} P_\nu(\psi_n = 1) \leqslant \limsup_{n\to\infty} m_n,$$

and since the above inequality holds for every ν, it also holds for the supremum i.e.

$$\sup_{\nu\in\mathcal{P}_0} \limsup_{n\to\infty} P_\nu(\psi_n = 1) \leqslant \limsup_{n\to\infty} m_n,$$

and since the right-hand side is less than α by assumption, we are done.

□

The motivation for defining the level of a test is to ensure that if we have a sufficiently large sample size, we can bound the probability that we reject H_0 falsely. Note however that the definition of pointwise asymptotic level yields that for any $\nu \in \mathcal{P}_0$ and any $\varepsilon > 0$, we can find $N \in \mathbb{N}$ such that for all $n \geqslant N$, we have $P_\nu(\psi_n = 1) \leqslant \alpha + \varepsilon$. In particular the choice of N is dependent on ν.

If we have uniformly asymptotic level, we get that for each $\varepsilon > 0$, there exists some $N \in \mathbb{N}$ such that the largest probability $P_\nu(\psi_n = 1)$ is less than $\varepsilon + \alpha$. In other words we can choose a threshold ε and then by working backwards, we can be sure that the actual level is within the threshold for all $\nu \in \mathcal{P}_0$ simultaneously.

The following is an example of a test that does not have uniform asymptotic level.

Example 2.2.7 (Pointwise vs. uniform asymptotic level)**.** This example is based on a similar example in [17]. Consider the statistical model consisting of all distributions on $\mathbb{R}$ with finite variance and the hypothesis that the distribution has mean zero. We consider the sequence of tests given by

$$\psi_n(x) = \begin{cases} 1 & \text{if } \frac{\bar{x}\sqrt{n}}{\hat{\sigma}} > z_{1-\alpha} \\ 0 & \text{otherwise} \end{cases},$$

where $\bar{x}$ is the empirical mean of x, $\hat{\sigma}$ is the unbiased estimate of the standard deviation of x and $z_{1-\alpha}$ is the $1-\alpha$ quantile of the normal distribution. Note for any $\nu \in \mathcal{P}_0$

$$P_\nu(\psi_n = 1) \to P(Z > z_{1-\alpha}) = \alpha,$$

as $n \to \infty$ since the central limit theorem yields that $\frac{\bar{X}\sqrt{n}}{\hat{\sigma}} \to Z$ as $n \to \infty$ where $Z \sim \mathcal{N}(0,1)$. Since the convergence holds for every ν it also holds for the supremum and thus

$$\sup_{\nu \in \mathcal{P}_0} \limsup_{n \to \infty} P_\nu(\psi_n = 1) = \alpha,$$

proving that the sequence of tests has pointwise asymptotic level.

This test does *not* achieve uniform asymptotic level since for any n and any $c \in (0,1)$, we can find a distribution $\nu \in \mathcal{P}_0$ such that $P_\nu(\psi_n = 1) \geqslant c$. To do this consider the distribution that puts mass $1-p$ on p and mass p on $-(1-p)$. Clearly this distribution has mean zero and finite variance so it is in $\mathcal{P}_0$.

Note that if given a sample x of size n from this distribution, the probability that all $x_i = p$ is $(1-p)^n$. If given such an observation, $\hat{\sigma} = 0$ and $\bar{x}$ is positive so $\psi_n(x) = 1$. This implies that $P_\nu(\psi_n = 1) \geqslant (1-p)^n$ and choosing $p = 1 - c^{1/n}$ shows that $P_\nu(\psi_n = 1) \geqslant c$ thus we do not have uniform asymptotic level.

Having defined tests we can turn to the problem of constructing them. A common strategy is to transform the sequence of observations in some way that has the same distribution for all $\nu \in \mathcal{P}$. This leads to the definition of a test statistic.

Definition 2.2.8 (Test statistics)**.** Let $(\mathcal{X}, \mathbb{E}, \mathcal{P})$ be a statistical model and $(g_n)_{n\in\mathbb{N}}$ be a sequence of functions where $g_n : \mathcal{X}^n \to \mathbb{R}$. Let furthermore $\mathcal{P}_0 \subseteq \mathcal{P}$.

1. If for all $n \in \mathbb{N}$, $g_n(X^{(n)})$ has the same continuous distribution for all $\nu \in \mathcal{P}_0$, we say that $(g_n)_{n\in\mathbb{N}}$ is a *test statistic with respect to* $\mathcal{P}_0$.

2. If $g_n(X^{(n)})$ converges in distribution to the same continuous distribution for all $\nu \in \mathcal{P}_0$, we say that $(g_n)_{n\in\mathbb{N}}$ is an *asymptotic test statistic with respect to* $\mathcal{P}_0$.

3. If $g_n(X^{(n)})$ converges in distribution uniformly over $\mathcal{P}_0$ to the same continuous distribution, we say that $(g_n)_{n\in\mathbb{N}}$ is a *uniform asymptotic test statistic with respect to* $\mathcal{P}_0$.

Remark 2.2.9 (Continuity of test statistic distributions)**.** In the definition above we have assumed that the (limiting) distributions of the test statistics are continuous. There is no a priori reason for this but it simplifies many of the upcoming proofs and considerations. Continuity of the distributions allows us to always find sets of arbitrary probability and for us not to discern between open and closed sets. To the best of the authors knowledge, most of the following results still hold if this assumption was omitted (with modifications to account for the possibility of point masses).

Using test statistics as defined above we can create tests.

Definition 2.2.10 (Tests from test statistics)**.** Let $(\mathcal{X}, \mathbb{E}, \mathcal{P})$ be a statistical model and $(\psi_n)_{n\in\mathbb{N}}$ be a sequence of tests of a hypothesis H_0 (with associated null-set of probability measures $\mathcal{P}_0$). Let further $\alpha \in (0,1)$.

1. If $(g_n)_{n\in\mathbb{N}}$ is a test statistic with respect to $\mathcal{P}_0$, we can find a sequence of sets $(\mathcal{B}_n)_{n\in\mathbb{N}}$ such that $P(g_n(X^{(n)}) \in \mathcal{B}_n) = \alpha$. From this we define a sequence of tests $(\psi_n)_{n\in\mathbb{N}}$ by

$$\psi_n(x^{(n)}) = \begin{cases} 1 & \text{if } g_n(x^{(n)}) \in \mathcal{B}_n \\ 0 & \text{otherwise} \end{cases}.$$

   Any such sequence of tests is called a *test constructed from the test statistic* $(g_n)_{n\in\mathbb{N}}$ *of level* α.

2. If $(g_n)_{n\in\mathbb{N}}$ is a (uniform) asymptotic test statistic with respect to $\mathcal{P}_0$, we can find a set $\mathcal{B}$ such that $P(V \in \mathcal{B}) = \alpha$ where V is a random variable with the same distribution as the limiting distribution of the asymptotic test statistic. We can then define a sequence of tests $(\psi_n)_{n\in\mathbb{N}}$ by

$$\psi_n(x^{(n)}) = \begin{cases} 1 & \text{if } g_n(x^{(n)}) \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}.$$

   Any such sequence of tests is called a *test constructed from the (uniform) asymptotic test statistic* $(g_n)_{n\in\mathbb{N}}$ *of level* α.

Most well-known statistical tests are constructed in one of the ways described above. This way of constructing tests allows us to immediately deduce various properties of the resulting tests.

Theorem 2.2.11 (Properties of tests from test statistics)**.** Let $(\mathcal{X}, \mathbb{E}, \mathcal{P})$ be a statistical model, $\alpha \in (0,1)$ and $(\psi_n)_{n\in\mathbb{N}}$ be a sequence of tests of a hypothesis H_0 (with associated null-set of probability measures $\mathcal{P}_0$) of level α.

1. If $(\psi_n)_{n\in\mathbb{N}}$ is constructed from a test statistic then $(\psi_n)_{n\in\mathbb{N}}$ has valid level.

2. If $(\psi_n)_{n\in\mathbb{N}}$ is constructed from an asymptotic test statistic then $(\psi_n)_{n\in\mathbb{N}}$ has pointwise asymptotic level.

3. If $(\psi_n)_{n\in\mathbb{N}}$ is constructed from a uniform asymptotic test statistic then $(\psi_n)_{n\in\mathbb{N}}$ has pointwise asymptotic level.

Proof.

1. Let $(g_n)_{n\in\mathbb{N}}$ denote the test statistic. Note that for every $\nu \in \mathcal{P}_0$ and $n \in \mathbb{N}$

$$P_\nu(\psi_n = 1) = P_\nu(g_n(X^n) \in \mathcal{B}_n) = \alpha,$$

   so it also holds for the sup over $\nu \in \mathcal{P}_0$, proving that the test has valid level.

2. Let $(g_n)_{n\in\mathbb{N}}$ denote the asymptotic test statistic. Note that for every $\nu \in \mathcal{P}_0$ we have

$$\limsup_{n\to\infty} P_\nu(\psi_n = 1) = \limsup_{n\to\infty} P_\nu(g_n(X^n) \in \mathcal{B})$$

   Now the convergence $g_n(X^n, \nu) \xrightarrow{\mathcal{D}} V$ yields immediately that

$$P_\nu(g_n(X^n, \nu) \in \mathcal{B}) \to P(V \in \mathcal{B}) = \alpha$$

   as $n \to \infty$. Now by taking sup over $\nu \in \mathcal{P}_0$ the result follows since sup respects inequalities.

3. Follows by arguments similar to the above by using the stronger assumption of uniform convergence in distribution of the test statistic.

□

We will apply this way of constructing tests when we construct the GCM in the following section.

2.3 Univariate GCM

In this section we define the Generalised Covariance Measure and prove its asymptotic properties under various assumptions.

To motivate the construction of the GCM, recall the statement of Theorem 2.1.15; if X and Y are real-valued random variables that are conditionally independent given a third random variable Z, then the product of the residuals of X and Y when regressing on Z will have mean zero. When given n observations of (X, Y, Z) we can perform the regression and calculate empirical versions of these residuals. We can then test whether the mean of the product of these residuals is zero to get a test of conditional independence. Let us be slightly more formal and start to define the quantities needed to prove the upcoming statements about the asymptotic level of the proposed test.

Definition 2.3.1 (Generalised Covariance Measure)**.** Let X and Y be univariate real-valued random variables and let Z be a random variable with values in the measurable space $(\mathcal{Z}, \mathcal{G})$. Consider the statistical model for X, Y and Z that contains all joint distributions on $\mathbb{R}^2 \times \mathcal{Z}$ i.e.

$$\mathcal{P} = \{\nu \text{ probability measure on } (\mathbb{R}^2 \times \mathcal{Z}, \mathbb{B}^2 \otimes \mathcal{G})\}.$$

Consider the hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$ with corresponding subset of probability measures $\mathcal{P}_0$. For every $\nu \in \mathcal{P}$, we can write

$$X = \underbrace{E_\nu(X \mid Z)}_{f_\nu(Z)} + \underbrace{X - E_\nu(X \mid Z)}_{\varepsilon_\nu},$$

i.e. $f_\nu(z) = E_\nu(X \mid Z = z)$ and similarly

$$Y = \underbrace{E_\nu(Y \mid Z)}_{g_\nu(Z)} + \underbrace{Y - E_\nu(Y \mid Z)}_{\xi_\nu}.$$

Let $(x, y, z)^n \in (\mathbb{R}^2 \times \mathcal{Z})^n$ be a sample of size n from the model and let $\widehat{f}^{(n)}$ and $\widehat{g}^{(n)}$ denote estimates of f and g based on the sample. For $i \in \{1, \ldots, n\}$ define

$$R_i^{(n)} = (x_i - \widehat{f}^{(n)}(z_i))(y_i - \widehat{g}^{(n)}(z_i))$$

and define

$$T_n = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n R_i^{(n)}}{\left(\frac{1}{n} \sum_{i=1}^n (R_i^{(n)})^2 - \left(\frac{1}{n} \sum_{j=1}^n R_j^{(n)} \right)^2 \right)^{\frac{1}{2}}}.$$

For $\alpha \in (0, 1)$ the sequence of tests $(\psi_n)_{n \in \mathbb{N}}$ given by

$$\psi_n((x, y, z)^{(n)}) = \begin{cases} 1 & \text{if } |T_n| > z_{1-\frac{\alpha}{2}} \\ 0 & \text{otherwise} \end{cases}$$

is the *Generalised Covariance Measure* with level α, where $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution.

Theorem 2.3.2 (GCM has asymptotic pointwise level)**.** Continuing from Definition 2.3.1, we define for each $\nu \in \mathcal{P}$

$$u_\nu(z) = E_\nu(\varepsilon_\nu^2 \mid Z = z), \quad v_\nu(z) = E_\nu(\xi_\nu^2 \mid Z = z)$$

We further define the mean-squared prediction error and weighted mean-squared prediction error for f

$$M_{\nu,n}^f = \frac{1}{n} \sum_{i=1}^n (f_\nu(z_i) - \widehat{f}^{(n)}(z_i))^2 \quad \text{and} \quad \tilde{M}_{\nu,n}^f = \frac{1}{n} \sum_{i=1}^n (f_\nu(z_i) - \widehat{f}^{(n)}(z_i))^2 v_\nu(z_i)$$

and

$$M^g_{\nu,n} = \frac{1}{n}\sum_{i=1}^{n}(g_\nu(z_i) - \widehat{g}^{(n)}(z_i))^2 \quad \text{and} \quad \tilde{M}^g_{\nu,n} = \frac{1}{n}\sum_{i=1}^{n}(g_\nu(z_i) - \widehat{g}^{(n)}(z_i))^2 u_\nu(z_i)$$

for g.

Assume that for each $\nu \in \mathcal{P}_0$, $nM^f_{\nu,n}M^g_{\nu,n} \xrightarrow{P} 0$, $\tilde{M}^f_{\nu,n} \xrightarrow{P} 0$, $\tilde{M}^g_{\nu,n} \xrightarrow{P} 0$ and $0 < E_\nu(\varepsilon_\nu^2\xi_\nu^2) < \infty$ then the GCM has pointwise asymptotic level.

Proof.
We show that for each $\nu \in \mathcal{P}_0$, T_n is an asymptotic test statistic, since Theorem 2.2.11 then implies that the GCM has pointwise asymptotic level. We will show that $T_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. To that end let $\nu \in \mathcal{P}_0$ be given and fixed. We will at times lighten notation and omit ν subscripts from expectations, probabilities and other expressions. Define

$$\tau_n^N = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} R_i^{(n)} \quad \text{and} \quad \tau_n^D = \left(\frac{1}{n}\sum_{i=1}^{n}\left(R_i^{(n)}\right)^2 - \left(\frac{1}{n}\sum_{j=1}^{n} R_j^{(n)}\right)^2\right)^{\frac{1}{2}},$$

so that T_n is the ratio of τ_n^N and τ_n^D. If we can show that $\tau_n^D \xrightarrow{P} \sqrt{\mathrm{Var}(\varepsilon\xi)}$ and $\tau_n^N \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{Var}(\varepsilon\xi))$ by Slutsky's theorem, we will be done. Note that we can decompose τ_n^N in the following way

$$\begin{aligned}
\tau_n^N &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(x_i) + \varepsilon_i - \widehat{f}^{(n)}(z_i))(g_\nu(y_i) + \xi_i - \widehat{g}^{(n)}(z_i)) \\
&= \underbrace{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\xi_i}_{U_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(z_i) - \widehat{f}^{(n)}(z_i))(g(z_i) - \widehat{g}^{(n)}(z_i))}_{a_n} \\
&\quad + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(z_i) - \widehat{f}^{(n)}(z_i))\xi_i}_{b_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(z_i) - \widehat{g}^{(n)}(z_i))\varepsilon_i}_{c_n}.
\end{aligned}$$

Note that by Theorem 2.1.15 the sequence $(\varepsilon_i\xi_i)_{i\in\mathbb{N}}$ has mean zero, since $X \perp\!\!\!\perp Y \mid Z$ for $\nu \in \mathcal{P}_0$, and by assumption it has finite variance (equal to the second moment of the sequence) so the CLT gives that $U_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, E(\varepsilon^2\xi^2))$. Cauchy-Schwarz inequality yields that

$$|a_n| \leqslant \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(f(z_i) - \widehat{f}^{(n)}(z_i))^2\sum_{i=1}^{n}(g(z_i) - \widehat{g}^{(n)}(z_i))^2} = \sqrt{nM_n^fM_n^g} \xrightarrow{P} 0,$$

since we have assumed that $nM_n^fM_n^g \xrightarrow{P} 0$. To show that $b_n \xrightarrow{P} 0$, we note that if $b_n^2 \xrightarrow{P} 0$ so does b_n. By Lemma 2.1.28 if we can show that $E(b_n^2 \mid (X_i)_{1\leqslant i\leqslant n}, (Z_i)_{1\leqslant i\leqslant n}) \xrightarrow{P} 0$ we will

thus have shown that $b_n \xrightarrow{P} 0$. Letting $X^{(n)} = (X_i)_{1 \leqslant i \leqslant n}$ and similarly $Z^{(n)}$, note that

$$E(b_n^2 \mid X^{(n)}, Z^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (f(z_i) - \widehat{f}^{(n)}(z_i))(f(z_j) - \widehat{f}^{(n)}(z_j)) E(\xi_i \xi_j \mid X^{(n)}, Z^{(n)}),$$

by the fact that the terms involving f and $\widehat{f}$ are measurable when knowing $X^{(n)}$ and $Z^{(n)}$. Since $\xi_i \xi_j$ only depends on Z_i and Z_j of the conditioning variables, we can drop the remaining variables from the conditioning expression. For $i \neq j$, by using that $E(Y_i \mid Z_i) = E(Y_i \mid Z_i, Z_j)$ since Z_j is independent of (Y_i, Z_i) and by pulling out what is known, we get

$$\begin{aligned} E(\xi_i \xi_j \mid X^{(n)}, Z^{(n)}) &= E(Y_i Y_j - E(Y_j \mid Z_j) Y_i - Y_j E(Y_i \mid Z_i) + E(Y_j \mid Z_j) E(Y_i \mid Z_i) \mid Z_i, Z_j) \\ &= E(Y_i Y_j \mid Z_i, Z_j) - E(Y_j \mid Z_i, Z_j) E(Y_i \mid Z_i, Z_j) = \operatorname{Cov}(Y_i, Y_j \mid Z_i, Z_j). \end{aligned}$$

By Theorem 2.1.14 this is zero if $Y_i \perp\!\!\!\perp Y_j \mid Z_i, Z_j$. By assumption we have $(Y_i, Z_i) \perp\!\!\!\perp (Y_j, Z_j)$ so applying weak union and symmetry from Theorem 2.1.9 yields the desired conditional independence statement. This lets us write

$$\begin{aligned} E(b_n^2 \mid X^{(n)}, Z^{(n)}) &= \frac{1}{n} \sum_{i=1}^{n} (f(z_i) - \widehat{f}^{(n)}(z_i))^2 E(\xi_i^2 \mid Z_i) \\ &= \tilde{M}_n^f \xrightarrow{P} 0 \end{aligned}$$

by assumption.

An analogous argument to the one above applies to c_n, thus $c_n \xrightarrow{P} 0$ and Slutsky's theorem yields that $\tau_n^N \xrightarrow{\mathcal{D}} \mathcal{N}(0, \operatorname{Var}(\varepsilon\xi))$. We now turn to τ_n^D and note that

$$(\tau_n^D)^2 = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(R_i^{(n)}\right)^2}_{p_n} - \underbrace{\left(\frac{1}{n} \sum_{j=1}^{n} R_j^{(n)}\right)^2}_{q_n}.$$

From the results above, we can easily conclude that $q_n \xrightarrow{P} 0$ since

$$q_n = \left(\frac{1}{\sqrt{n}} \tau_n^N\right)^2,$$

and τ_n^N converges in distribution to a normal distribution, while $\frac{1}{\sqrt{n}}$ converges to 0 in probability, so Slutsky's theorem yields that their product converges in distribution to 0. Convergence in distribution to a constant is equivalent to convergence in probability to the same constant and squaring retains convergence by the continuous mapping theorem, proving that $q_n \xrightarrow{P} 0$.

We now intend to show that $p_n \xrightarrow{P} \mathrm{Var}(\varepsilon_\nu \xi_\nu)$. We decompose p_n into 9 terms as seen below

$$\begin{aligned}
p_n &= \frac{1}{n}\sum_{i=1}^{n}(f(z_i)+\varepsilon_i-\widehat{f}^{(n)}(z_i))^2(g(z_i)+\xi_i-\widehat{g}^{(n)}(z_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[\Big((f(z_i)-\widehat{f}^{(n)}(z_i))^2+\varepsilon_i^2+2(f(z_i)-\widehat{f}^{(n)}(z_i))\varepsilon_i\Big) \\
&\cdot\Big((g(z_i)-\widehat{g}^{(n)}(z_i))^2+\xi_i^2+2(g(z_i)-\widehat{g}^{(n)}(z_i))\xi_i\Big)\Big] \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\xi_i^2}_{\mathrm{I}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(f(z_i)-\widehat{f}^{(n)}(z_i))^2(g(z_i)-\widehat{g}^{(n)}(z_i))^2}_{\mathrm{II}_n} \\
&+ \underbrace{\frac{4}{n}\sum_{i=1}^{n}(f(z_i)-\widehat{f}^{(n)}(z_i))(g(z_i)-\widehat{g}^{(n)}(z_i))\varepsilon_i\xi_i}_{\mathrm{III}_n} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}(f(z_i)-\widehat{f}^{(n)}(z_i))^2\xi_i^2}_{\mathrm{IV}_n^f} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(g(z_i)-\widehat{g}^{(n)}(z_i))^2\varepsilon_i^2}_{\mathrm{IV}_n^g} \\
&+ \underbrace{\frac{2}{n}\sum_{i=1}^{n}(f(z_i)-\widehat{f}^{(n)}(z_i))^2(g(z_i)-\widehat{g}^{(n)}(z_i))\xi_i}_{\mathrm{V}_n^f} + \underbrace{\frac{2}{n}\sum_{i=1}^{n}(f(z_i)-\widehat{f}^{(n)}(z_i))\varepsilon_i\xi_i^2}_{\mathrm{VI}_n^f} \\
&+ \underbrace{\frac{2}{n}\sum_{i=1}^{n}(g(z_i)-\widehat{g}^{(n)}(z_i))^2(f(z_i)-\widehat{f}^{(n)}(z_i))\varepsilon_i}_{\mathrm{V}_n^g} + \underbrace{\frac{2}{n}\sum_{i=1}^{n}(g(z_i)-\widehat{g}^{(n)}(z_i))\xi_i\varepsilon_i^2}_{\mathrm{VI}_n^g}.
\end{aligned}$$

By the Law of Large Numbers, we have

$$\mathrm{I}_n \xrightarrow{P} E(\varepsilon^2\xi^2),$$

and as noted earlier $\varepsilon\xi$ has mean zero, so this is also the variance of $\varepsilon\xi$. Note that for positive sequences a_n and b_n, we have $\sum_i a_i b_i \leqslant \sum a_i \sum b_i$ (easily seen by noting that every term on the LHS appears on the RHS), from which we can get

$$\mathrm{II}_n \leqslant n M_n^f M_n^g \xrightarrow{P} 0$$

by assumption. By Cauchy-Schwarz inequality we get

$$\mathrm{III}_n \leqslant 4\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\xi_i^2\right)^{\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^{n}(f(z_i)-\widehat{f}^{(n)}(z_i))^2(g(z_i)-\widehat{g}^{(n)}(z_i))^2\right)^{\frac{1}{2}} = 4\mathrm{I}_n^{\frac{1}{2}}\mathrm{II}_n^{\frac{1}{2}} \xrightarrow{P} 0,$$

since we just showed that I_n is convergent and II_n goes to 0 in probability, their product goes to 0 in probability and the continuous mapping theorem yields that the same is true when taking square roots.

By Lemma 2.1.28 $\mathrm{IV}_n^f \xrightarrow{P} 0$ since

$$E(\mathrm{IV}_n^f \mid X^{(n)}, Z^{(n)}) = \frac{1}{n}\sum_{i=1}^{n}(f(z_i) - \widehat{f}^{(n)}(z_i))^2 E(\xi_i^2 \mid X^{(n)}, Z^{(n)}) = \tilde{M}_n^f \xrightarrow{P} 0$$

by assumption and similarly for IV_n^g. By the triangle inequality and Cauchy-Schwarz, we have

$$\begin{aligned}
|\mathrm{V}_n^f| &\leqslant \frac{2}{n}\sum_{i=1}^{n}|g(z_i) - \widehat{g}^{(n)}(z_i)||f(z_i) - \widehat{f}^{(n)}(z_i)||g(z_i) - \widehat{g}^{(n)}(z_i)||\xi_i| \\
&\leqslant 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(z_i) - \widehat{f}^{(n)}(z_i))^2(g(z_i) - \widehat{g}^{(n)}(z_i))^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(g(z_i) - \widehat{g}^{(n)}(z_i))^2\xi_i^2} \\
&= 2\sqrt{\mathrm{II}_n \mathrm{IV}_n^g} \xrightarrow{P} 0
\end{aligned}$$

by the results above and similarly for V_n^g. Finally by using the triangle inequality and Cauchy-Schwarz again, we have

$$\begin{aligned}
|\mathrm{VI}_n^f| &\leqslant \frac{2}{n}\sum_{i=1}^{n}|f(z_i) - \widehat{f}^{(n)}(z_i)||\xi_i||\varepsilon_i||\xi_i| \\
&\leqslant 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(z_i) - \widehat{f}^{(n)}(z_i))^2\xi_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\xi_i^2} \\
&= 2\sqrt{\mathrm{IV}_n^g \mathrm{I}_n} \xrightarrow{P} 0
\end{aligned}$$

by the results above and similarly for VI_n^g.

This shows that $p_n \xrightarrow{P} \mathrm{Var}(\varepsilon\xi)$ thus by the continuous mapping theorem $\tau_n^D \xrightarrow{P} \sqrt{\mathrm{Var}(\varepsilon\xi)}$ and by Slutsky's theorem $T_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$ as desired. □

Using the uniform convergence results given in the section on probabilistic preliminaries, we can also argue for the uniform asymptotic level of the GCM under stronger assumptions.

Theorem 2.3.3 (GCM has uniform asymptotic level)**.** Consider the same setup as in Theorem 2.3.2 Let $\tilde{\mathcal{P}}_0 \subseteq \mathcal{P}_0$. Assume that for each $\nu \in \tilde{\mathcal{P}}_0$, $nM_{\nu,n}^f M_{\nu,n}^g \xrightarrow[\tilde{\mathcal{P}}_0]{P} 0$, $\tilde{M}_{\nu,n}^f \xrightarrow[\tilde{\mathcal{P}}_0]{P} 0$, $\tilde{M}_{\nu,n}^g \xrightarrow[\tilde{\mathcal{P}}_0]{P} 0$, $0 < \inf_{\nu\in\tilde{\mathcal{P}}_0} E_\nu(\varepsilon_\nu^2\xi_\nu^2)$ and $\sup_{\nu\in\tilde{\mathcal{P}}_0} E_\nu(\varepsilon_\nu^{2+\eta}\xi_\nu^{2+\eta}) < \infty$ for some $\eta > 0$, then the GCM has uniform asymptotic level wrt. $\tilde{\mathcal{P}}_0$.

Proof.
We can repeat the arguments in Theorem 2.3.2 with the stronger assumptions, applying the uniform versions of the continuous mapping theorem, Slutsky's theorem, the law of large numbers and the central limit thoerem (Theorem 2.1.18, Theorem 2.1.23, Theorem 2.1.25 and Theorem 2.1.26 respectively). We note that we do rely on the unproven Assumption 2.1.24. □

We have now proven that the test holds asymptotic level under checkable assumptions. One should however note that there are examples of X and Y not being conditionally independent given Z but where the mean of the product of the residuals is still zero (one such example is where X and Y are independent Rademacher and $Z = XY$). For such distributions the GCM would always accept the null hypothesis of conditional independence despite this being false.

Although X and Y are assumed to be univariate in the theorem above, it is possible to generalize the GCM to the multivariate setting in several ways. A straightforward generalization would be to simply proceed as above but instead letting R_i be the outer product of the residuals. One could then construct a test statistic that was asymptotically standard normal of dimension equal to the product of the dimensions of X and Y under the null. The norm of such a test statistic is chi-squared with degrees of freedom equal to the dimension of the normal distribution. In the original article by Shah & Peters they propose instead considering each combination of components of X and Y, calculating the one-dimensional test statistic and aggregating by taking the maximum. Both strategies lead to valid tests and in the original article it is argued that the maximum-based test has a smaller bias.

One worthwhile thing to notice is the immediate lack of assumptions on Z. In theory Z could take values in any measurable space. Of course we still need to be able to regress X and Y on Z in practice and have results about the mean squared error of such a process for the result to hold. With Z being uni- or multivariate and real-valued the applications are obvious but one could also consider Z to be a functional random variable. This would mean letting Z take values in a Hilbert or Banach space of functions. To illustrate this point we will include the following theorem from [26] about the convergence rate of a functional linear model. The article by Shin & Lee discusses a model where predictors are both functional and multivariate but we will simplify and only give the result for the functional linear model. The functional definitions mentioned in the theorem are omitted for brevity but reading the following two chapters should provide most of the required background to understand the theorem.

Theorem 2.3.4 (Mean squared prediction error in functional linear model)**.** Let Z be a functional random variable defined on $[0, 1]$, i.e. a random variable in $L^2([0, 1])$ with finite

second moment and covariance operator $\mathscr{K}$ and let ε be a real-valued random variable independent from Z with $E(\varepsilon) = 0$ and $E(\varepsilon^2) = \sigma^2$. Let γ be a function in $L^2([0,1])$ and define the random variable Y by

$$Y = \langle Z, \gamma \rangle + \varepsilon := \int_0^1 \gamma(t) Z(t) \, \mathrm{d}t + \varepsilon.$$

This is the *functional linear model with scalar response*.

Let $(Y_i, Z_i)_{i \in \mathbb{N}}$ be an i.i.d. sequence of realizations generated by the model. We can then estimate γ consistently using a principal components method as described in [26] yielding an estimate $\widehat{\gamma}$. Assume that Z has finite fourth moment and let $(\lambda_j, \phi_j)_{j \in \mathbb{N}}$ denote the eigenvalue and -vector pairs of the covariance operator $\mathscr{K}$. Assume that

1. There exists $C_1 > 0$ so that for all $j \geqslant 1$, we have $E(\langle X, \phi_j \rangle^4) \leqslant C_1 \lambda_j^2$.
2. There exists $C_2 > 0$ and $a > 1$ so that for all $j \geqslant 1$, we have $C^{-1} j^{-a} \leqslant \lambda_j \leqslant C j^{-a}$ and $\lambda_j - \lambda_{j+1} \geqslant C j^{-a-1}$.
3. There exists $C_3 > 0$ and $b > 1/2$ such that for all $j \geqslant 1$, we have $|\langle \gamma, \phi_j \rangle| \leqslant C j^{-b}$.

Then for a new independent observation $\tilde{Z}$, we have

$$\sqrt{n} E\left((\langle \widehat{\gamma}, \tilde{Z} \rangle - \langle \gamma, \tilde{Z} \rangle)^2 \mid (Y_i, Z_i)_{1 \leqslant i \leqslant n} \right) \xrightarrow{P} 0.$$

Under some technical smoothness conditions, we do in fact achieve a mean square prediction error that is sufficient for the requirements given in Theorem 2.3.2. While the linear relationship required is a rather strong condition more scalar-on-function regression methods are actively being developed (see [20] for an overview of methods.)

If we consider a situation where the functional linear model is applicable, we now have a concrete example of a conditional independence test with pointwise asymptotic level. To the best of the authors knowledge, this is a novel result and the first example of a conditional independence test involving functional data.

Theorem 2.3.5 (GCM in the functional linear model with scalar response)**.** Let X and Y be univariate random variables and let Z be a functional random variable defined on $[0,1]$ as in Theorem 2.3.4. Assume that both (X, Z) and (Y, Z) satisfy the conditions in Theorem 2.3.4 and that furthermore both $u_\nu(z)$ and $v_\nu(z)$ in Theorem 2.3.2 are bounded by some $\sigma^2 > 0$ for all ν. Then the GCM has asymptotic pointwise level when testing whether $X \perp\!\!\!\perp Y \mid Z$.

Proof.
We will only need to show that $\sqrt{n} M^f_{\nu,n}$ and $\sqrt{n} M^g_{\nu,n}$ go to zero in probability, since the

remaining conditions then follow by the assumptions. This holds by Markov's inequality and Theorem 2.3.4 . □

There are many technical assumptions in the theorem above but note that most of them will hold if Z is a functional Gaussian. If a more general regression method was applied, we would probably be able to drop many of the technical assumptions required in Theorem 2.3.4. Theorem 2.3.5 allows us to test for conditional independence when X and Y are univariate real and Z is functional. Ideally we would like to consider X, Y and Z all being functional in nature and testing conditional independence. The rest of the thesis will be devoted to the pursuit of generalizing the GCM to the case where X, Y and Z belong to a Hilbert space, which encapsulates both data types. In the upcoming chapter we will delve into the theory of Hilbert spaces.

Hilbert spaces, operator theory and Bochner integration

In this chapter we will give an overview of the theory of Hilbert spaces, which will form the foundation of the subsequent development of probability and statistics on Hilbert spaces. We will review some of the fundamental properties of Hilbert spaces and then proceed to define linear functionals and operators between Hilbert spaces. Finally we will show how to construct an integral for functions with values in a Hilbert space. Throughout this chapter we will be sparse with proofs in an attempt to be reasonably brief and due to the well-known nature of many of the given results. For the full proofs about the geometry and fundamentals of Hilbert spaces see [23] or [12] for proofs about operators and integrals on Hilbert spaces.

3.1 Fundamental properties of Hilbert spaces

In this section we motivate and give the fundamental definitions and theorems regarding Hilbert spaces. These are mainly results about decompositions of variables in the space or the space itself using the inner product.

Recall the usual "nice" properties of the Euclidean spaces $\mathbb{R}^d$: we have a well-defined distance measure, a size of each element (a norm), a sense of orthogonality through an inner product and the space has "no holes" in the sense that if we have a Cauchy sequence in $\mathbb{R}^d$, we can find a limit of the sequence in $\mathbb{R}^d$ i.e. the space is complete. Hilbert spaces are a generalization of the Euclidean spaces that retain all of these concepts (and thus $\mathbb{R}^d$ are all Hilbert spaces) but also include more abstract spaces that can be of infinite dimension. We define a Hilbert space below (for a review of some essential definitions of topology, algebra and analysis, see the appendix).

Definition 3.1.1 (Hilbert space)**.** A *Hilbert space* is an inner product space over $\mathbb{R}$ or $\mathbb{C}$ that is complete with respect to the metric induced by the inner product.

Throughout this thesis we will concentrate solely on Hilbert spaces over $\mathbb{R}$. As previously mentioned the Euclidean spaces are all Hilbert spaces but lets consider some more exotic examples.

Example 3.1.2 (ℓ^2)**.** Let $\mathbb{R}^{\mathbb{N}}$ denote the set of all sequences with values in $\mathbb{R}$ and denote by $\ell^2(\mathbb{N})$ (or simply ℓ^2 for short) the subset of $\mathbb{R}^{\mathbb{N}}$ of square-summable sequences i.e.

$$\ell^2(\mathbb{N}) = \left\{ x \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \sum_{n=1}^{\infty} x_n^2 < \infty \right\}.$$

It is straight-forward to show that ℓ^2 is an inner product space with inner product for $x, y \in \ell^2$

$$\langle x, y \rangle = \sum_{n=1}^{\infty} x_n y_n,$$

which is finite by Cauchy-Schwarz inequality. The norm is given by

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{n=1}^{\infty} x_n^2}.$$

Using the tools of real analysis, one can show that ℓ^2 is in fact complete and is thus a Hilbert space [23].

Example 3.1.3 ($L^2[0,1]$)**.** Let $([0,1], \mathbb{B}_{[0,1]}, m_{[0,1]})$ denote the unit interval with the Borel σ-algebra restricted to the unit interval and the Lebesgue measure m. Consider the set of measurable functions from $[0,1]$ to $\mathbb{R}$ denoted by $\mathcal{M}[0,1]$. Let $\mathcal{L}^2[0,1]$ be the subset of $\mathcal{M}[0,1]$ given by

$$\mathcal{L}^2[0,1] = \left\{ f \in \mathcal{M}[0,1] \;\middle|\; \int_{[0,1]} f^2 \, \mathrm{d}m_{[0,1]} < \infty \right\}.$$

Define an equivalence relation on $\mathcal{L}^2[0,1]$ such that $f \sim g \iff m(f \neq g) = 0$ and construct the quotient space $L^2[0,1]$ consisting of the equivalence classes under the aforementioned relation. We will typically abuse notation slightly and still refer to the elements of $L^2[0,1]$ as functions despite them being equivalence classes. It is immediate that this is an inner product space with inner product for $f, g \in L^2[0,1]$

$$\langle f, g \rangle = \int_{[0,1]} fg \, \mathrm{d}m_{[0,1]},$$

which is finite by Hölder's inequality. The norm is given by

$$\|f\| = \sqrt{\langle x, x \rangle} = \sqrt{\int_{[0,1]} f^2 \, \mathrm{d}m_{[0,1]}}.$$

The completeness of $L^2[0,1]$ is a deep result in analysis: the Riesz-Fisher theorem (see [23] for a proof) and using the conclusions of that theorem, we get that $L^2[0,1]$ is a Hilbert space.

A favoured property of the Euclidean spaces is the existence of a basis: a linearly independent set such that every element of the vector space can be written as a linear combination of basis elements. We can define a similar concept for Hilbert spaces:

Definition 3.1.4 (Orthonormality, orthogonality and ONB's)**.** Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$.

$x, y \in \mathcal{H}$ are said to be *orthogonal* if $\langle x, y \rangle = 0$.

A set of elements $\{e_i\}_{i \in I}$, where I is some index sex, is said to be *orthonormal* if $\|e_n\| = 1$ for all $n \in \mathbb{N}$ and if the elements of the sequence are pairwise orthogonal.

If $\text{span}(\{e_i\}_{i \in I})$ is also dense in $\mathcal{H}$, it is said to be an *orthonormal basis* (ONB) for $\mathcal{H}$. The *dimension* of a Hilbert space is the cardinality of I.

Note that the definition above is not identical to the linear algebra definition of a basis. We require that every element of $\mathcal{H}$ can be approximated arbitrarily well with linear combinations of basis elements whereas the linear algebra definition of a basis requires the existence of a linear combination equalling the element. The different definitions are however identical for finite-dimensional spaces. Having an ONB allows us to express elements of the Hilbert space using "coordinates" and finding these coordinates can be done using the inner product as can be seen from the following result:

Theorem 3.1.5 (Fourier expansion and Parseval's identity)**.** Every element x of a Hilbert space $\mathcal{H}$ with ONB $\{e_i\}_{i \in I}$ can be written

$$x = \sum_{i \in I} \langle x, e_i \rangle e_i,$$

which is called the *Fourier expansion* of x and furthermore we have Parseval identity:

$$\|x\|^2 = \sum_{i \in I} \langle x, e_i \rangle^2.$$

We can also express the inner product of two elements $x, y \in \mathcal{H}$ as

$$\langle x, y \rangle = \sum_{i \in I} \langle x, e_i \rangle \langle y, e_i \rangle.$$

For a proof of Theorem 3.1.5 and the following theorems, see [23]. The usefulness of a basis diminishes greatly if the index set is not countable. This is at least partly due to the fact that for uncountable sums to be finite, only a countable number of terms can be non-zero.

Theorem 3.1.6 gives a characterization of when a Hilbert space has a countable ONB. Recall that a space is separable if it contains a countable, dense subset.

Theorem 3.1.6 (Separability and countable ONB's)**.** A Hilbert space is separable if and only if it has a countable orthonormal basis.

Let us see some examples of ONB's for the previous examples.

Example 3.1.7 (ONB for ℓ^2)**.** Consider ℓ^2 as in Example 3.1.2. Consider the ℓ^2 elements e_i which are sequences with a 1 in the i'th position and zero elsewhere. The sequence $(e_i)_{i=1}^{\infty}$ is an ONB for ℓ^2. Note that the index set is $\mathbb{N}$ so ℓ^2 is infinite-dimensional. The ONB is countable, so ℓ^2 is separable.

Example 3.1.8 (ONB for $L^2[0,1]$)**.** Consider again $L^2[0,1]$ as defined in Example 3.1.3. Consider the sets of $L^2[0,1]$ elements

$$B_1 = \{f_n(x) = \sqrt{2}\sin(n\pi x)\}$$
$$B_2 = \{f_0(x) = 1\} \cup \{f_n(x) = \sqrt{2}\cos(n\pi x) \mid n \in \mathbb{N}\}$$
$$B_3 = \{f_0(x) = 1\} \cup \{f_{2n-1}(x) = \sqrt{2}\sin(2n\pi x) \mid n \in \mathbb{N}\} \cup \{f_{2n}(x) = \sqrt{2}\cos(2n\pi x) \mid n \in \mathbb{N}\}.$$

These are all examples of ONB's for $L^2[0,1]$ (see [12] Theorem 2.4.18 for a proof of this fact). Note that all the bases are indexed by $\mathbb{N}_0$ or $\mathbb{N}$, thus the space is infinite-dimensional and separable.

In many fields of mathematics we identify two spaces as being "almost the same" (or more formally isomorphic or congruent) if the structure of the spaces is the same even if the objects have different names. There is a notion of congruence of metric spaces:

Definition 3.1.9 (Isomorphic metric spaces)**.** Two metric spaces (M_1, d_1) and (M_2, d_2) are said to be *isomorphic* or *congruent* if there exists a bijective function $\Psi : M_2 \to M_1$ such that

$$d_2(x_1, x_2) = d_1(\Psi(x_1), \Psi(x_2)) \quad \forall x_1, x_2 \in M_2.$$

We're often only interested in spaces modulo congruence, since if the two spaces are congruent, they have the same "structure". With that in mind we can note the following result.

Theorem 3.1.10. Every infinite-dimensional separable Hilbert space is congruent to ℓ^2.

This shows that we can essentially think of any separable infinite-dimensional Hilbert space as ℓ^2. This space is the canonical choice for a separable infinite-dimensional Hilbert space due to the many well-known results about summation of sequences and due to the obvious choice of ONB in the space.

3.2 Operators on Hilbert spaces

In this section we will give some results about linear mappings between Hilbert spaces (operators) and functionals on Hilbert spaces. This theory is crucial for the later development of covariances of Hilbertian random variables and for Hilbertian linear models.

Having defined the fundamental properties of a single Hilbert space, we can start considering what happens when we consider mappings between and on Hilbert spaces. We will concentrate on the mappings that preserve the underlying vector space structure; the homomorphisms. For vector spaces these are exactly the linear maps and for finite-dimensional spaces we have a fully developed theory of linear algebra to understand these mappings. We would also like to preserve the topological structure of the spaces, so we further restrict ourselves to continuous mappings. On a finite-dimensional space every linear mapping is continuous but this is not the case on infinite-dimensional spaces. However for linear maps continuity is intimately connected to boundedness as we will now show.

Definition 3.2.1 (Bounded and linear operators)**.** Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be normed vector spaces with norms $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively. Let furthermore $\mathscr{A} : \mathcal{X}_1 \to \mathcal{X}_2$ be a mapping.

We say that $\mathscr{A}$ is *linear* if $\mathscr{A}(cx) = c\mathscr{A}(x)$ and if $\mathscr{A}(x+y) = \mathscr{A}(x) + \mathscr{A}(y)$ for all $x \in \mathcal{X}_1$ and $c \in \mathbb{R}$.

We say that a linear mapping $\mathscr{A}$ is *bounded* if there exists $C > 0$ such that $\|\mathscr{A}x\|_2 \leqslant C\|x\|_1$ for all $x \in \mathcal{X}_1$.

Theorem 3.2.2 (Linear operators are uniformly continuous iff they are bounded)**.** Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be normed vector spaces. Let furthermore $\mathscr{A} : \mathcal{X}_1 \to \mathcal{X}_2$ be a linear mapping. Then $\mathscr{A}$ is bounded if and only if it is uniformly continuous.

Proof.
Let $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the norms of $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively.

If $\mathscr{A}$ is uniformly continuous, it is in particular continuous at 0, so we can find $\delta > 0$, so $\|\mathscr{A}x\|_2 < 1$ for all $x \in \mathcal{X}_1$ such that $\|x\|_1 < \delta$. Then using linearity of $\mathscr{A}$ and the norm, we get

$$\|\mathscr{A}x\|_2 = \left\|\mathscr{A}\left(\frac{\delta x}{\|x\|_1}\right)\right\|_2 \frac{\|x\|_1}{\delta} \leqslant \frac{1}{\delta}\|x\|_1,$$

which proves the first implication. For the converse note that boundedness trivially implies that the functional is Lipschitz, which implies that it is uniformly continuous. □

This leads to the following fundamental definition of a bounded linear operator and the space of bounded linear operators.

Definition 3.2.3 (Space of bounded linear operators)**.** Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be normed vector spaces. We denote by $\mathfrak{B}(\mathcal{X}_1, \mathcal{X}_2)$ the set of all bounded linear mappings from $\mathcal{X}_1$ to $\mathcal{X}_2$. The elements of $\mathfrak{B}(\mathcal{X}_1, \mathcal{X}_2)$ are called *bounded linear operators* or simply *operators*. If $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$ we write $\mathfrak{B}(\mathcal{X})$.

By convention for $\mathscr{A} \in \mathfrak{B}(\mathcal{X}_1, \mathcal{X}_2)$ we often write $\mathscr{A}x$ to denote $\mathscr{A}(x)$ for $x \in \mathcal{X}_1$.

We define the *rank* of an operator $\mathscr{A} \in \mathfrak{B}(\mathcal{X}_1, \mathcal{X}_2)$ by

$$\operatorname{rank}(\mathscr{A}) = \dim(\operatorname{Im}(\mathscr{A})).$$

Note that the rank of an operator be infinite. The space of bounded linear operators has some particularly nice properties, for instance it is quite easy to see that it is a vector space. Furthermore we can define a norm and show that if the codomain of the operators is complete, then so is the space of bounded linear operators. Recall that a complete normed space is called a Banach space.

Theorem 3.2.4 (Operator norm and completeness of bounded linear operators)**.** Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be normed vector spaces with norms $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively and consider the space of bounded linear operators $\mathfrak{B}(\mathcal{X}_1, \mathcal{X}_2)$. For $\mathscr{A} \in \mathfrak{B}(\mathcal{X}_1, \mathcal{X}_2)$ we define the *operator norm* of $\mathscr{A}$ as

$$\|\mathscr{A}\| = \sup_{x \in \mathcal{X}_1, \|x\|_1 = 1} \|\mathscr{A}x\|_2.$$

For any $x \in \mathcal{X}_1$, we have the fundamental inequality

$$\|\mathscr{A}x\|_2 \leqslant \|\mathscr{A}\| \|x\|_1,$$

and if $\mathcal{X}_2$ is a Banach space then so is $\mathfrak{B}(\mathcal{X}_1, \mathcal{X}_2)$ under the operator norm.

With these results these results in mind, we proceed to focus on results more specific to Hilbert spaces. We will mainly consider two cases, the bounded linear functionals on a Hilbert space and bounded linear operators between Hilbert spaces. We start by considering the functionals and introduce the notion of a dual space.

Definition 3.2.5 (Dual space)**.** Let $\mathcal{X}$ be a normed vector space, we define the *dual space* of $\mathcal{X}$ as $\mathfrak{B}(\mathcal{X}, \mathbb{R})$ and denote it by $\mathcal{X}^*$.

One motivation for introducing the idea of a dual space is the fact that we often understand a space by understanding the well-behaved functions that act on it. One of the surprising facts about the dual space of a Hilbert space is Riesz representation theorem, which is proved in [23].

Theorem 3.2.6 (Riesz representation theorem)**.** Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|_{\mathcal{H}}$ and let $\mathscr{A} \in \mathcal{H}^*$. Then there exists a unique element $h_{\mathscr{A}} \in \mathcal{H}$ called the *representer* of $\mathscr{A}$, with the property that

$$\mathscr{A}h = \langle h, h_{\mathscr{A}} \rangle, \quad \forall h \in \mathcal{H}$$

and $\|\mathscr{A}\| = \|h_{\mathscr{A}}\|_{\mathcal{H}}$ where $\|\cdot\|$ denotes the operator norm.

Note that the Riesz representation theorem implies that $\mathcal{H}$ is self-dual, i.e. the dual space of $\mathcal{H}$ is congruent with $\mathcal{H}$.

Corollary 3.2.7 (Hilbert spaces are self-dual)**.** Let $\mathcal{H}$ be a Hilbert space. Then $\mathcal{H}^*$ is congruent to $\mathcal{H}$.

We thus have a complete characterization of the linear functionals and their behaviour, since they simply rely on the properties of the inner product. This will prove invaluable for many proofs later.

In the remainder of this section we will consider the bounded linear operators between two Hilbert spaces. We can think of such operators as generalizations of operators between Euclidean spaces which we would typically represent as matrices. We start by proving a relationship between $\mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ and $\mathfrak{B}(\mathcal{H}_2, \mathcal{H}_1)$ through a generalization of transposition of matrices.

Theorem 3.2.8 (Adjoint operators and their existence)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ respectively. For every $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ there exists a unique element $\mathscr{A}^* \in \mathfrak{B}(\mathcal{H}_2, \mathcal{H}_1)$ such that

$$\langle \mathscr{A}h_1, h_2 \rangle_2 = \langle h_1, \mathscr{A}^* h_2 \rangle_1, \quad \forall h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2.$$

Proof.
For each $h_2 \in \mathcal{H}_2$, consider the bounded linear functional $\phi \in \mathcal{H}_1^*$, given by

$$\phi(h_1) = \langle \mathscr{A}h_1, h_2 \rangle_2.$$

Linearity is obvious and boundedness follows from Cauchy-Schwarz. By Riesz representation theorem there exists a unique representer $y \in \mathcal{H}_1$, such that

$$\phi(h_1) = \langle h_1, y \rangle_1.$$

This lets us define for each $h_2 \in \mathcal{H}_2$, $\mathscr{A}^* h_2 := y$, and we thus have

$$\langle \mathscr{A}h_1, h_2 \rangle_2 = \langle h_1, \mathscr{A}^* h_2 \rangle_1,$$

as desired. It remains to show that $\mathscr{A}^*$ is a bounded linear operator. Linearity can be seen by noting that the ϕ's for linear combinations of elements in $\mathcal{H}_2$ decompose linearly. To get boundedness, we note that (letting $\|\cdot\|$ denote the operator norm)

$$\|\mathscr{A}^* h_2\|_1^2 = \langle \mathscr{A}^* h_2, \mathscr{A} h_2 \rangle_1 = \langle \mathscr{A}\mathscr{A}^* h_2, h_2 \rangle_2 \leqslant \|\mathscr{A}\mathscr{A}^* h_2\|_2 \|h_2\|_2 \leqslant \|\mathscr{A}\| \|\mathscr{A}^* h_2\|_1 \|h_2\|_2,$$

where we have written the norm using the inner product, applied that $\mathscr{A}^*$ is the adjoint of $\mathscr{A}$, applied Cauchy-Schwarz and used the operator norm inequality. Dividing through by $\|\mathscr{A}^* h_2\|_1$ proves boundedness and therefore also the result. □

This leads to the following definition.

Definition 3.2.9 (Self-adjoint operators)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. Let $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$.

The operator defined as $\mathscr{A}^*$ in Theorem 3.2.8 is called the *adjoint* of $\mathscr{A}$. If $\mathcal{H}_1 = \mathcal{H}_2$ and $\mathscr{A} = \mathscr{A}^*$ then $\mathscr{A}$ is said to be *self-adjoint*.

The adjoint of an operator and the operator itself share several nice properties as the following theorem shows (see [12] for a proof of these properties).

Proposition 3.2.10 (Properties of adjoint operators)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be separable Hilbert spaces and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$. Then

1. $(\mathscr{A}^*)^* = \mathscr{A}$.
2. $\|\mathscr{A}^*\| = \|\mathscr{A}\|$.
3. $\mathscr{A}^*\mathscr{A}$ and $\mathscr{A}\mathscr{A}^*$ are self-adjoint.
4. $\|\mathscr{A}^*\mathscr{A}\| = \|\mathscr{A}\|^2$.

Recall that these are all properties of matrix transposition and matrix transposition is exactly the finite-dimensional version of finding an adjoint operator. For operators in $\mathfrak{B}(\mathcal{H})$ we have further special properties defined below.

Definition 3.2.11 (Definiteness of operators)**.** Let $\mathcal{H}$ be a Hilbert space and consider $\mathscr{A} \in \mathfrak{B}(\mathcal{H})$.

We say that $\mathscr{A}$ is *non-negative definite* or simply *non-negative* if it is self-adjoint and if

$$\langle \mathscr{A} h, h \rangle \geqslant 0, \quad \forall h \in \mathcal{H}.$$

We say that $\mathscr{A}$ is *positive definite* or *positive* if the inequality is strict.

Non-negative operators are nice because they admit a square-root decomposition as the following theorem illustrates.

Proposition 3.2.12 (Square roots of non-negative definite operators)**.** Let $\mathcal{H}$ be a Hilbert space and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H})$ be a non-negative operator. Then there exists a unique operator $\mathscr{S} \in \mathfrak{B}(\mathcal{H})$ such that $\mathscr{S}^2 = \mathscr{A}$ and such that $\mathscr{S}$ commutes with any operator that commutes with $\mathscr{A}$. We call this operator the *square-root operator* of $\mathscr{A}$ and denote it by $\mathscr{A}^{1/2}$.

When working with an operator $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$, we could sometimes be interested in the inverse operation. For the Euclidean spaces we are used to inverting matrices as long as they are bijective. This still applies in general, i.e. if an operator is bijective ($\mathrm{Ker}(\mathscr{A}) = \{0\}$ and $\mathrm{Im}(\mathscr{A}) = \mathcal{H}_2$) we know that an inverse exists. The question is whether it is bounded which the following theorem asserts in the affirmative.

Proposition 3.2.13 (Inverse operators)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces and consider $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$. If $\mathscr{A}$ is bijective and $\mathscr{A}^{-1}$ is the inverse operator, then $\mathscr{A}^{-1} \in \mathfrak{B}(\mathcal{H}_2, \mathcal{H}_1)$.

Unfortunately we shall see later that the operators we see in practice on Hilbert spaces are rarely bijective and thus the existence of an inverse operator can barely ever be assumed. The operators we see in practice are the compact operators and let us immediately define what it means for an operator to be compact.

Definition 3.2.14 (Compact operator)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. A linear mapping $\mathscr{A} : \mathcal{H}_1 \to \mathcal{H}_2$ is said to be *compact* if for every bounded sequence $\{x_n\}_{n=1}^{\infty}$, the sequence $\{\mathscr{A}x_n\}_{n=1}^{\infty}$ contains a convergent subsequence.

Compact operators behave much like the well-known linear transformations on the Euclidean spaces, that can be represented by matrices. Let us note a few properties of compact operators.

Theorem 3.2.15 (Properties of compact operators)**.** Let $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ be Hilbert spaces and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ and $\mathscr{B} \in \mathfrak{B}(\mathcal{H}_2, \mathcal{H}_3)$. Then

1. If $\mathrm{rank}(\mathscr{A}) < \infty$ then $\mathscr{A}$ is compact.
2. If $\mathscr{A}$ or $\mathscr{B}$ is compact, then so is $\mathscr{B}\mathscr{A}$.
3. $\mathscr{A}$ is compact if and only if there exists a sequence of operators $\{\mathscr{A}_n\}_{n=1}^{\infty}$ with finite rank such that $\|\mathscr{A} - \mathscr{A}_n\| \to 0$ as $n \to \infty$ where $\|\cdot\|$ denotes the operator norm.
4. $\mathscr{A}$ is compact if and only if $\mathscr{A}^*$ is compact.

Proof.
We only prove the first two claims and refer to [12] for a proof of the remaining two claims.

For the first claim note that if $\mathrm{Im}(\mathscr{A})$ is finite-dimensional, it is isomorphic to $\mathbb{R}^n$ and thus the result follows immediately from Bolzano-Weierstrass theorem.

For the second claim, let $(h_n)_{n\in\mathbb{N}}$ be a bounded sequence in $\mathcal{H}_1$. If $\mathscr{A}$ is compact, the sequence $(\mathscr{A}h_n)_{n\in\mathbb{N}}$ contains a convergent subsequence and since $\mathscr{B}$ is continuous, $(\mathscr{B}\mathscr{A}h_n)_{n\in\mathbb{N}}$ also contains a convergent subsequence and thus $\mathscr{A}\mathscr{B}$ is compact. If instead $\mathscr{B}$ is compact, we note that the sequence $(\mathscr{A}h_n)_{n\in\mathbb{N}}$ is bounded, since $\mathscr{A}$ is continuous, and therefore $(\mathscr{B}\mathscr{A}h_n)_{n\in\mathbb{N}}$ contains a convergent subsequence, since $\mathscr{B}$ is compact, which proves that $\mathscr{B}\mathscr{A}$ is compact. □

The theorem above shows that the compact operators are exactly the generalization of the finite-rank operators in the sense that they can be approximated arbitrarily well by finite-rank operators. Unlike the usual finite-dimensional cases the identity operator is not compact as we shall see below.

Theorem 3.2.16 (Identity operator is not compact)**.** Let $\mathcal{H}$ be an infinite-dimensional Hilbert space. Then the identity operator on $\mathcal{H}$ is not compact.

Proof.
Let $\|\cdot\|$ denote the norm on $\mathcal{H}$, $\mathscr{I}$ denote the identity operator and let $(e_n)_{n\in\mathbb{N}}$ be an orthonormal basis for $\mathcal{H}$. Note that $(e_n)_{n\in\mathbb{N}}$ is a bounded sequence and that for $i \neq j$

$$\|\mathscr{I}e_i - \mathscr{I}e_j\| = \|e_i - e_j\| = \sqrt{\|e_i\|^2 + \|e_j\|^2} = \sqrt{2},$$

by Parseval's identity. Therefore $(\mathscr{I}e_n)_{n\in\mathbb{N}}$ does not contain a convergent subsequence and thus the operator is not compact. □

It is perhaps a little surprising that the identity operator is not well-behaved and this will have consequences for the probability theory we shall develop later. As alluded to earlier, we can show that none of the the compact operators are bijective and thus we cannot invert them.

Corollary 3.2.17 (Bijective operators are not compact)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be infinite-dimensional Hilbert spaces and assume that $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ is bijective. Then $\mathscr{A}$ is not compact.

Proof.
Note that

$$\mathscr{A}^{-1}\mathscr{A} = \mathscr{I},$$

where $\mathscr{I}$ denotes the identity operator on $\mathcal{H}_1$. If either $\mathscr{A}$ or its inverse were compact, this would contradict Theorem 3.2.15 since $\mathscr{I}$ is not compact and therefore both $\mathscr{A}$ and $\mathscr{A}^{-1}$ are not compact. □

The compact operators are nice in particular because of the spectral theory developed for them. Recall that for the Euclidean spaces, when given a symmetric matrix, we can find an eigen-decomposition of the matrix into an orthonormal basis. This is the crucial theoretical underpinnings of principal component analysis for multivariate random variables amongst other powerful results. We would like to generalize the concept of eigen-decompositions and compact operators are exactly the operators where this is possible as we shall see. Let us define eigenvectors and -values.

Definition 3.2.18 (Eigen-decomposition of operators)**.** Let $\mathcal{H}$ be a Hilbert space and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H})$. Assume that there exists $\lambda \in \mathbb{R}$ and $e \in \mathcal{H}$ so that

$$\mathscr{A}e = \lambda e,$$

then we say that λ is an *eigenvalue* of $\mathscr{A}$ and e is an *eigenvector* (or sometimes *eigenfunction* if $\mathcal{H}$ is a function space).

This is basically the same definition as for matrices and finite-dimensional linear transformations. We will at times draw upon the notion of an outer product on Hilbert spaces which generalizes the outer product of vectors on the Euclidean spaces.

Definition 3.2.19 (Outer product)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$. Let $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$. The linear mapping from $\mathcal{H}_1$ to $\mathcal{H}_2$ given by

$$h_1 \odot_1 h_2 = \langle \cdot, h_1 \rangle_1 h_2$$

is called the *outer product of* h_1 *with* h_2. Similarly we define the outer product of h_2 with h_1 as

$$h_2 \odot_2 h_1 = \langle \cdot, h_2 \rangle_2 h_1.$$

If $\mathcal{H}_1 = \mathcal{H}_2$ we simply write $h_1 \odot h_2$ and $h_2 \odot h_1$ respectively.

Remark 3.2.20 (Outer product notation)**.** The outer product is typically denoted by $\otimes$, since it is intimately connected with the tensor product of Hilbert spaces. We will not delve into the theory of tensor products and to avoid any confusion for other readers unfamiliar with this theory, we will instead employ the symbol $\odot$ to denote the outer product.

Some authors also define the outer product $h_1 \odot_1 h_2 = \langle \cdot, h_2 \rangle_2 h_1$ i.e. the opposite of what was done in this thesis. We follow the notation in [12] and as such have chosen to do it as in the definition.

Let us note some simple properties of the outer product.

Theorem 3.2.21 (Properties of the outer product)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_1$ and $\langle\cdot,\cdot\rangle_2$ and norms $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively. Let $h_1, \tilde{h}_1 \in \mathcal{H}_1$ and $h_2, \tilde{h}_2 \in \mathcal{H}_2$ and $a, b \in \mathbb{R}$. Then

1. $h_1 \odot_1 h_2 \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ with $\|h_1 \odot_1 h_2\| = \|h_1\|_1\|h_2\|_2$, where $\|\cdot\|$ denotes the operator norm.
2. $(h_1 + \tilde{h}_1) \odot_1 (h_2 + \tilde{h}_2) = h_1 \odot_1 h_2 + h_1 \odot_1 \tilde{h}_2 + \tilde{h}_1 \odot_1 h_2 + \tilde{h}_1 \odot_1 \tilde{h}_2$.
3. $(ah_1) \odot_1 (bh_2) = ab(h_1 \odot_1 h_2)$.
4. $\operatorname{rank}(h_1 \odot_1 h_2) = 1$ if h_1 and h_2 are non-zero.
5. $(h_1 \odot_1 h_2)^* = h_2 \odot_2 h_1$.

Proof.

1. By definition and Cauchy-Schwarz, we have

$$\|h_1 \odot_1 h_2\| = \sup_{\tilde{h}\in\mathcal{H}_1, \|\tilde{h}\|_1} \|\langle\tilde{h}, h_1\rangle_1 h_2\|_2 \leqslant \sup_{\tilde{h}\in\mathcal{H}_1, \|\tilde{h}\|_1} \|\tilde{h}\|_1\|h_1\|_1\|h_2\|_2 = \|h_1\|_1\|h_2\|_2.$$

   Setting $\tilde{h} = \frac{h_1}{\|h_1\|}$ yields equality in the above, thus proving the statement.

2. Consider how the operator acts on a $h \in \mathcal{H}_1$ and use properties of the inner product.
3. Similar to above.
4. Every element in $\operatorname{Im}(h_1 \odot_1 h_2)$ can be written on the form $c \cdot h_2$ for some $c \in \mathbb{R}$. Conversely, given some $c \in \mathbb{R}$, we can find $h \in \mathcal{H}_1$, so that $\langle h, h_1\rangle = c$, thus $\operatorname{Im}(h_1 \odot_1 h_2) = \operatorname{span}(h_2)$, which is one-dimensional and therefore $\operatorname{rank}(h_1 \odot_1 h_2) = 1$.
5. Straightforward calculations show that for any $\tilde{h}_1 \in \mathcal{H}_1$ and $\tilde{h}_2 \in \mathcal{H}_2$, we have

$$\begin{aligned}\langle(h_1 \odot_1 h_2)\tilde{h}_1, \tilde{h}_2\rangle_2 &= \langle\langle h_1, \tilde{h}_1\rangle_1 h_2, \tilde{h}_2\rangle_2 = \langle h_1, \tilde{h}_1\rangle_1\langle h_2, \tilde{h}_2\rangle_2 \\ &= \langle\tilde{h}_1, \langle\tilde{h}_2, h_2\rangle_2 h_1\rangle_1 = \langle\tilde{h}_1, (h_2 \odot_2 h_1)\tilde{h}_2\rangle_2,\end{aligned}$$

   proving the statement.

□

With that in mind, we can give the spectral theorem for compact, self-adjoint operators.

Theorem 3.2.22 (Spectral theorem for compact and self-adjoint operators)**.** Let $\mathcal{H}$ be a Hilbert space and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H})$ be a compact, self-adjoint operator.

The set of non-zero eigenvalues of $\mathscr{A}$ is either finite or consists of a sequence tending to zero. Each non-zero eigenvalue has finite multiplicity and eigenvectors corresponding to different eigenvalues are orthogonal.

Letting $|\lambda_1| \geqslant |\lambda_2| \geqslant \dots$ denote the eigenvalues of $\mathscr{A}$ and $(e_n)_{n=1}^{\infty}$ denote the corresponding eigenvectors, we can apply the Gram-Schmidt procedure to $(e_n)_{n=1}^{\infty}$ and get an orthonormal basis $(\tilde{e}_n)_{n=1}^{\infty}$ of $\overline{\mathrm{Im}(\mathscr{A})}$ such that

$$\mathscr{A} = \sum_{j=1}^{\infty} \lambda_j \tilde{e}_j \odot \tilde{e}_j,$$

i.e. for every $h \in \mathcal{H}$

$$\mathscr{A}h = \sum_{j=1}^{\infty} \lambda_j \langle \tilde{e}_j, h \rangle \tilde{e}_j.$$

If $\mathscr{A}$ is also non-negative-definite, then all the eigenvalues are non-negative.

Ideally we would like a way to decompose any compact operator between two Hilbert spaces. Recall that for any operator $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ the operator $\mathscr{A}^*\mathscr{A}$ was self-adjoint. If $\mathscr{A}$ is compact, we know that composition with a bounded operator produces a compact operator, thus $\mathscr{A}^*\mathscr{A}$ is also compact. This implies that $\mathscr{A}^*\mathscr{A}$ has an eigen-decomposition and similarly for $\mathscr{A}\mathscr{A}^*$. This leads to the following theorem and definition of the singular value decomposition for operators that is a straightforward generalization of the concept for matrices.

Theorem 3.2.23. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ be a compact operator. Then denoting $(\lambda_j^2)_{j=1}^{\infty}$ the non-ascending eigenvalues of $\mathscr{A}^*\mathscr{A}$ (or equivalently $\mathscr{A}\mathscr{A}^*$), $(e_{1j})_{j=1}^{\infty}$ the orthonormal eigenvectors of $\mathscr{A}^*\mathscr{A}$ and $(e_{2j})_{j=1}^{\infty}$ the orthonormal eigenvectors of $\mathscr{A}\mathscr{A}^*$ such that $\mathscr{A}^* e_{2j} = \lambda_j e_{1j}$, we have

$$\mathscr{A} = \sum_{j=1}^{\infty} \lambda_j e_{1j} \odot_1 e_{2j},$$

i.e. for $h \in \mathcal{H}_1$

$$\mathscr{A}h = \sum_{j=1}^{\infty} \lambda_j \langle e_{1j}, h \rangle_1 e_{2j},$$

where $\langle \cdot, \cdot \rangle_1$ denotes the inner product in $\mathcal{H}_1$. We call the decomposition above the *singular value decomposition* and $((\lambda_j^2, e_{1j}, e_{2j}))_{j=1}^{\infty}$ a *singular system* for $\mathscr{A}$.

The singular value decomposition of a compact operator has many applications and uses, for instance it lets us calculate the norm of an operator with ease.

Theorem 3.2.24 (Operator norm is largest singular value)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ be a compact operator with largest singular value λ_1. Then

$$\|\mathscr{A}\| = \lambda_1.$$

The singular value decomposition of a compact operator is in fact a fundamental property of compact operators as the next theorem shows.

Theorem 3.2.25 (Compact if and only if singular value decomposition)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$. $\mathscr{A}$ is compact if and only if $\mathscr{A}$ has a singular value decomposition.

Proof.
We know that every compact operator has a singular value decomposition, so to prove the converse, assume that we're given an operator $\mathscr{A}$ with singular system $((\lambda_j^2, e_{1j}, e_{2j}))_{j=1}^{\infty}$ and define

$$\mathscr{A}_n = \sum_{j=1}^{n} \lambda_j (e_{1j} \odot_1 e_{2j}).$$

Each $\mathscr{A}_n$ is obviously finite-dimensional and we have $\|\mathscr{A} - \mathscr{A}\| = \lambda_{n+1}$, by Theorem 3.2.24, which goes to 0 as $n \to \infty$. We have now approximated $\mathscr{A}$ by a sequence of finite-dimensional operators so by Theorem 3.2.15, $\mathscr{A}$ is also compact, proving the statement. □

Now that we have defined and explored the compact operators, we will define the final two classes of operators, that we will need for the later work; the Hilbert-Schmidt operators and the trace class operators. We start by considering the Hilbert-Schmidt operators.

Definition 3.2.26 (Hilbert-Schmidt operators)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces, let $(e_i)_{i \in I}$ be an orthonormal basis for $\mathcal{H}_1$ and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$. Letting $\|\cdot\|_2$ denote the norm on $\mathcal{H}_2$, if

$$\sum_{i \in I} \|\mathscr{A} e_i\|_2^2 < \infty,$$

then $\mathscr{A}$ is called a *Hilbert-Schmidt operator*. The collection of Hilbert-Schmidt operators in $\mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ is denoted $\mathfrak{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$. The space of Hilbert-Schmidt operators is an inner-product space with

$$\langle \mathscr{A}, \mathscr{B} \rangle_{HS} = \sum_{i \in I} \langle \mathscr{A} e_i, \mathscr{B} e_i \rangle_2$$

for $\mathscr{A}, \mathscr{B} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$, where $\langle \cdot, \cdot \rangle_2$ is the inner product on $\mathcal{H}_2$. The corresponding norm is

$$\|\mathscr{A}\|_{HS}^2 = \sum_{i \in I} \|\mathscr{A} e_i\|_2^2.$$

For a proof that the construction above is well-defined, see [12].

Let us note some nice properties of Hilbert-Schmidt operators and in particular how the outer product behaves with the Hilbert-Schmidt norm and inner product.

Theorem 3.2.27 (Properties of Hilbert-Schmidt operators)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_1$ and $\langle\cdot,\cdot\rangle_2$ and norms $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively. Let $\mathscr{A} \in \mathfrak{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$ and further $h_1, \tilde{h}_1 \in \mathcal{H}_1$ and $h_2, \tilde{h}_2 \in \mathcal{H}_2$.

1. $\mathscr{A}^* \in \mathfrak{B}_{HS}(\mathcal{H}_2, \mathcal{H}_1)$.
2. $\mathscr{A}$ is compact.
3. If $(\lambda_j)_{j=1}^{\infty}$ are the singular values for $\mathscr{A}$ then
$$\|\mathscr{A}\|_{HS}^2 = \sum_{j=1}^{\infty} \lambda_j^2.$$
4. $\langle h_1 \odot_1 h_2, \tilde{h}_1 \odot_1 \tilde{h}_2\rangle_{HS} = \langle h_1, \tilde{h}_1\rangle_1 \langle h_2, \tilde{h}_2\rangle_2$.
5. $\|h_1 \odot_1 h_2\|_{HS} = \|h_1\|_1 \|h_2\|_2$.

Proof.
For proofs of the first three properties, see [12].

For the fourth claim, let $(e_i)_{i\in I}$ be an ONB for $\mathcal{H}_1$. We get by definition

$$\begin{aligned}\langle h_1 \odot_1 h_2, \tilde{h}_1 \odot_1 \tilde{h}_2\rangle_{HS} &= \sum_{i\in I} \langle\langle e_i, h_1\rangle_1 h_2, \langle e_i, \tilde{h}_1\rangle_1 \tilde{h}_2\rangle_2 \\ &= \langle h_2, \tilde{h}_2\rangle_2 \sum_{i\in I} \langle e_i, h_1\rangle_1 \langle e_i, \tilde{h}_1\rangle_1 = \langle h_1, \tilde{h}_1\rangle_1 \langle h_2, \tilde{h}_2\rangle_2,\end{aligned}$$

where the final equality is due to Parseval's identity.

The fifth claim follows immediately from the fourth. □

The fact that the Hilbert-Schmidt norm can be written as the sum of the squared singular values is highly useful in practice. The space of Hilbert-Schmidt operators is particularly nice because it forms a Hilbert space and we can explicitly construct an ONB for the space by combining ONB's of the domain and image Hilbert spaces.

Theorem 3.2.28 (Basis for Hilbert-Schmidt operators)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces with orthonormal bases $(e_{1i})_{i\in I}$ and $(e_{2j})_{j\in J}$ respectively. The space $\mathfrak{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$ is a Hilbert space and has an orthonormal basis given by $(e_{1i} \odot_1 e_{2j})_{(i,j)\in I\times J}$.

The theorem above also states, that if both Hilbert spaces are separable, the space of Hilbert-Schmidt operators will also be separable. Note also that the theorem shows that the finite rank operators are dense in the space of Hilbert-Schmidt operators since the orthonormal basis consists of rank one operators.

We saw before that the Hilbert-Schmidt operators have square-summable singular values, so a natural extension could be to consider operators with summable singular values. These are the final class of operators; the trace class operators.

Definition 3.2.29 (Trace-class operators)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces and let $(e_i)_{i\in I}$ be an orthonormal basis for $\mathcal{H}_1$. Denoting the inner product on $\mathcal{H}_1$ by $\langle\cdot,\cdot\rangle_1$, an operator $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ is said to be *trace-class* if

$$\|\mathscr{A}\|_{TR} := \sum_{i\in I} \langle (\mathscr{A}^*\mathscr{A})^{1/2} e_i, e_i\rangle_1$$

is finite. We call the quantity $\|\mathscr{A}\|_{TR}$ the *trace norm* of $\mathscr{A}$. We denote the space of all trace-class operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ by $\mathfrak{B}_{TR}(\mathcal{H}_1, \mathcal{H}_2)$.

An argument akin to the one employed for Hilbert-Schmidt operators could show that this is independent of the choice of orthonormal basis for $\mathcal{H}_1$. The trace class operators have several nice properties as seen below.

Theorem 3.2.30 (Properties of trace-class operators)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces and let $\mathscr{A} \in \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2)$ be trace class. Then

1. $\mathscr{A}$ is Hilbert-Schmidt and compact.

2. If $(\lambda_j)_{j=1}^{\infty}$ are the singular values of $\mathscr{A}$ then

$$\|\mathscr{A}\|_{TR} = \sum_{j=1}^{\infty} \lambda_j.$$

3. If $\mathcal{H}_1 = \mathcal{H}_2$ and $\mathscr{A}$ is self-adjoint with eigenvalue sequence denoted by $(\lambda_j)_{j=1}^{\infty}$ then

$$\|\mathscr{A}\|_{TR} = \sum_{j=1}^{\infty} |\lambda_j|.$$

We shall later see that the natural extension of the covariance of random variables to the Hilbert space setting, leads to the covariance being a trace class operator.

3.3 Integration of Hilbertian functions

In this section we will develop a rigorous theory of integration for functions with values in a separable Hilbert space. This is the theory of Bochner integration and constructs integrals in a way paralleling the construction of the Lebesgue integral for real-valued functions.

When defining the integral of a measurable real-valued function defined on some background measure space $(\mathcal{X}, \mathbb{E}, \mu)$, we consider an approximating sequence of simple functions where the integrals are obvious and define the integral of the function as the limit of integrals of simple functions. We will construct a similar idea for Hilbertian functions by first defining what it means for a function to be measurable. While most of the integration theory developed will work in more general spaces, we will present the theory solely for Hilbert spaces.

Recall that for an $\mathbb{R}$-valued function f, measurability amounts to the pre-image of every set $E \in \mathbb{E}$ under f to be an element of the Borel σ-algebra. Even though the concept of a Borel σ-algebra generalizes well, this concept of measurability is not always useful on general infinite-dimensional spaces. The key property that we would like to retain is the idea of approximating a function with simple functions. This is the notion of strong measurability as defined below.

Definition 3.3.1 (Simple Hilbertian function)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. A function $f : \mathcal{X} \to \mathcal{H}$ is said to be *simple* if there exists $k \in \mathbb{N}$, $A_1, \ldots, A_k \in \mathbb{E}$ and $h_1, \ldots, h_k \in \mathcal{H}$ such that

$$f(x) = \sum_{i=1}^{n} 1_{A_i}(x) h_i.$$

Note that the representation is not unique.

Definition 3.3.2 (Strong measurability of Hilbertian functions)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. A function $f : \mathcal{X} \to \mathcal{H}$ is said to be *strongly measurable* if there exists a sequence of simple functions $f_n : \mathcal{X} \to \mathcal{H}$ such that f_n converges to f pointwise, i.e. for each $x \in \mathcal{X}$ we have $\lim_{n \to \infty} f_n(x) = f(x)$.

As mentioned previously, we could also generalize the notion of a Borel σ-algebra and thus define the more familiar Borel measurability. Recall that the Borel σ-algebra on a metric space, $\mathcal{X}$, denoted by $\mathbb{B}(\mathcal{X})$, is the smallest σ-algebra containing the open sets on the space.

Definition 3.3.3 (Borel measurability of Hilbertian functions)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. A function $f : \mathcal{X} \to \mathcal{H}$ is said to be *Borel measurable* if

$$\forall B \in \mathbb{B}(\mathcal{H}) : f^{-1}(B) \in \mathbb{E}.$$

Finally there is a third form of measurability known as weak measurability that transforms the concern of measurability to the well-known case of $\mathbb{R}$-valued functions through the linear functionals on $\mathcal{H}$ i.e. using the inner product.

Definition 3.3.4 (Weak measurability of Hilbertian functions)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. A function $f : \mathcal{X} \to \mathcal{H}$ is said to be *weakly measurable* if the function $\langle f, h \rangle$ is measurable as a real-valued function for all $h \in \mathcal{H}$.

The interplay between these forms of measurability is well-studied and the crucial result is the following theorem by Pettis.

Theorem 3.3.5 (Pettis measurability theorem)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. Let $f : \mathcal{X} \to \mathcal{H}$ be some function.

We say that the function is *separably valued* if there exists a separable closed subset S of $\mathcal{H}$ such that $f(x) \in S$ for all $x \in \mathcal{X}$.

The following are equivalent:

1. f is strongly measurable.
2. f is separably valued and weakly measurable.
3. f is separably valued and Borel measurable.

For a proof of the theorem, see [24]. The Pettis measurability theorem also holds for the μ-almost everywhere equivalent properties above.

Remark 3.3.6 (Measurability on separable Hilbert spaces)**.** When working in a separable Hilbert space every function is separably valued, so the three notions of measurability coincide. To avoid any measurability concerns we will henceforth assume that the Hilbert spaces we are working on are separable. Thus we will simply call a function *measurable* if it satisfies any of the three definitions and use them interchangeably.

With the measurability concerns out of the way, we will proceed to define the Bochner integral of a Hilbertian function by first defining integrability and the integral of a simple function.

Definition 3.3.7 (Bochner integrals and integrability of simple functions)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. Any simple function f with decomposition

$$f(x) = \sum_{i=1}^{k} 1_{A_i}(x) h_i$$

is said to be *integrable* if $\mu(A_i) < \infty$ for all $i \in \{1, \ldots, k\}$ and the *Bochner integral* of f is defined as

$$\int_{\mathcal{X}} f \, \mathrm{d}\mu = \sum_{i=1}^{k} h_i \mu(A_i).$$

We can extend this to a generic measurable function in the same way that this was done for the usual Lebesgue integration theory.

Definition 3.3.8 (Bochner integrability and integrals)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. Let furthermore $f : \mathcal{X} \to \mathcal{H}$ be measurable. We say that f is *Bochner integrable* if there exists a sequence $(f_n)_{n=1}^{\infty}$ of simple and integrable functions such that

$$\lim_{n \to \infty} \int_{\mathcal{X}} \|f_n - f\| \, \mathrm{d}\mu = 0.$$

In this case we define the *Bochner integral* of f as

$$\int_{\mathcal{X}} f \, \mathrm{d}\mu = \lim_{n \to \infty} \int_{\mathcal{X}} f_n \, \mathrm{d}\mu.$$

Let us prove that the above construction is well-defined in the sense that the integral of the simple functions does not depend on its representation and the integral of a measurable function does not depend on the specific approximating sequence of simple functions.

Theorem 3.3.9 (Bochner integrals are well-defined)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. The Bochner integral of both simple functions and general measurable functions from $\mathcal{X}$ to $\mathcal{H}$ is well-defined.

Proof.
The integral of simple functions is independent of the representation by the same arguments that are used for the Lebesgue integral, see for instance [23] Lemma 9.1.

We still need to prove that the limit in the definition of the Bochner integral exists for non-simple f and is independent of the choice of approximating sequence. To that end let $f : \mathcal{X} \to \mathcal{H}$ be Bochner integrable and let $(f_n)_{n=1}^{\infty}$ be an approximating sequence of simple functions.

We start by showing that the integrals of the simple functions form a Cauchy sequence. Note that for any simple function f_n, we have from the triangle inequality

$$\left\| \int_{\mathcal{X}} f_n \, \mathrm{d}\mu \right\| \leqslant \int_{\mathcal{X}} \|f_n\| \, \mathrm{d}\mu.$$

In particular this holds for $f_n - f_m$ for $m, n \in \mathbb{N}$ and therefore again by the triangle inequality

$$\left\| \int_{\mathcal{X}} f_n \, \mathrm{d}\mu - \int_{\mathcal{X}} f_m \, \mathrm{d}\mu \right\| \leqslant \int_{\mathcal{X}} \|f_n - f_m\| \, \mathrm{d}\mu \leqslant \int_{\mathcal{X}} \|f_n - f\| \, \mathrm{d}\mu + \int_{\mathcal{X}} \|f - f_m\| \, \mathrm{d}\mu,$$

which goes to zero by assumption. This shows that the integrals are a Cauchy sequence and since $\mathcal{H}$ is complete, the limit exists.

If $(g_n)_{n=1}^{\infty}$ was another approximating sequence of simple functions, we could represent both f_n and g_n using the same sets and the triangle inequality would thus still be applicable to the integral of their difference. We would get

$$\left\| \int_{\mathcal{X}} f_n \,\mathrm{d}\mu - \int_{\mathcal{X}} g_n \,\mathrm{d}\mu \right\| \leqslant \int_{\mathcal{X}} \|f_n - g_n\| \,\mathrm{d}\mu \leqslant \int_{\mathcal{X}} \|f_n - f\| \,\mathrm{d}\mu + \int_{\mathcal{X}} \|f - g_n\| \,\mathrm{d}\mu,$$

which again converges to zero, showing that the limit is the same for all approximating sequences. □

The criterion of integrability is rather unwieldy in practice but fortunately we have the following theorem, that gives an easier condition to check.

Theorem 3.3.10 (Hilbertian functions are integrable if their norm is integrable)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. Let $f : \mathcal{X} \to \mathcal{H}$ be a measurable function and assume that $\int_{\mathcal{X}} \|f\| \,\mathrm{d}\mu < \infty$. Then f is Bochner integrable.

For a proof of this, see [12]. The Bochner integral has all of the nice properties of the usual integral including dominated convergence and the triangle inequality:

Theorem 3.3.11 (Dominated convergence theorem for Bochner integral)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. Let $f_n : \mathcal{X} \to \mathcal{H}$ be a sequence of Bochner integrable functions that converges to some $f : \mathcal{X} \to \mathcal{H}$. If there exists a non-negative Lebesgue integrable function g such that $\|f_n\| \leqslant g$ for all n μ-a.e., then f is Bochner integrable and

$$\int_{\mathcal{X}} f \,\mathrm{d}\mu = \lim_{n \to \infty} \int_{\mathcal{X}} f_n \,\mathrm{d}\mu.$$

Theorem 3.3.12 (Triangle inequality for Bochner integral)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space. Let $f : \mathcal{X} \to \mathcal{H}$ be a Bochner integrable function. Then

$$\left\| \int_{\mathcal{X}} f \,\mathrm{d}\mu \right\| \leqslant \int_{\mathcal{X}} \|f\| \,\mathrm{d}\mu.$$

An application of dominated convergence yields that the Bochner integral is also well-behaved when working with sequences of integrable functions.

Theorem 3.3.13 (Interchanging series and Bochner integrals)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $\mathcal{H}$ be a Hilbert space with norm $\|\cdot\|$. Let $f_n : \mathcal{X} \to \mathcal{H}$ be a sequence of Bochner integrable functions such that

$$\int_{\mathcal{X}} \sum_{n=1}^{\infty} \|f_n\| \,\mathrm{d}\mu < \infty,$$

(by the usual theorems for Lebesgue integration, this is equivalent to $\sum_{n=1}^{\infty}\int_{\mathcal{X}}\|f_n\|\,\mathrm{d}\mu<\infty$) then $\sum_{n=1}^{\infty}f_n(x)$ converges μ-a.e. and

$$\sum_{n=1}^{\infty}\int_{\mathcal{X}}f_n\,\mathrm{d}\mu=\int_{\mathcal{X}}\sum_{n=1}^{\infty}f_n\,\mathrm{d}\mu.$$

Proof.
Define the partial sums $h_n(x)=\sum_{i=1}^{n}f_i(x)$, the full series $h(x)=\sum_{n=1}^{\infty}f_n(x)$ and the series of norms $g(x)=\sum_{i=1}^{\infty}\|f_n(x)\|$.

Note first that the assumption of integrability of g implies that g is finite μ-a.e. immediately. This in turn implies that h is finite μ a.e. and that $h_n\to h$ as $n\to\infty$.

Note further that by the triangle inequality

$$\|h_n(x)\|\leqslant\sum_{i=1}^{n}\|f_n(x)\|\leqslant\sum_{i=1}^{\infty}\|f_n(x)\|=g(x).$$

We can now apply dominated convergence; Theorem 3.3.11 and get that h is integrable and that

$$\int_{\mathcal{X}}h\,\mathrm{d}\mu=\lim_{n\to\infty}\int_{\mathcal{X}}h_n(x)\,\mathrm{d}\mu=\lim_{n\to\infty}\sum_{i=1}^{n}\int_{\mathcal{X}}f_i(x)\,\mathrm{d}\mu=\sum_{n=1}^{\infty}\int_{\mathcal{X}}f_n(x)\,\mathrm{d}\mu$$

as desired. □

One of the most desirable properties of the Bochner integral is the fact that it is well-behaved when composed with operators.

Theorem 3.3.14 (Interchanging operators and Bochner integrals)**.** Let $(\mathcal{X},\mathbb{E},\mu)$ be a measure space and let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. Let $\mathscr{A}\in\mathfrak{B}(\mathcal{H}_1,\mathcal{H}_2)$ and let $f:\mathcal{X}\to\mathcal{H}_1$ be a Bochner integrable function. Then $\mathscr{A}f$ is Bochner integrable and

$$\mathscr{A}\left(\int_{\mathcal{X}}f\,\mathrm{d}\mu\right)=\int_{\mathcal{X}}\mathscr{A}f\,\mathrm{d}\mu.$$

When we start working with covariances of Hilbertian random variables, we shall see many random operators, i.e. operator-valued random variables. The following theorem shows that Bochner integrals are also well-behaved for these mappings when the mapping takes values in the space of Hilbert-Schmidt operators.

Theorem 3.3.15 (Interchanging integrals and operator-valued mappings)**.** Let $(\mathcal{X},\mathbb{E},\mu)$ be a measure space and let $\mathcal{H}_1$ and $\mathcal{H}_2$ be separable Hilbert spaces. Let $\mathscr{F}:\mathcal{X}\to\mathfrak{B}_{HS}(\mathcal{H}_1,\mathcal{H}_2)$ be an operator-valued mapping and assume that it is Bochner integrable i.e. $\int_{\mathcal{X}}\|\mathscr{F}\|_{HS}\,\mathrm{d}\mu<\infty$. Then for any $h\in\mathcal{H}_1$

$$\int_{\mathcal{X}}\mathscr{F}h\,\mathrm{d}\mu=\left(\int_{\mathcal{X}}\mathscr{F}\,\mathrm{d}\mu\right)h.$$

Proof.
Let $h \in \mathcal{H}_1$ be given and define the mapping $\mathscr{G} : \mathfrak{B}(\mathcal{H}_1, \mathcal{H}_2) \to \mathcal{H}_2$ by $\mathscr{G}(\mathscr{F}) = \mathscr{F}h$. With this definition the desired result becomes

$$\int_{\mathcal{X}} \mathscr{G}(\mathscr{F}) \, \mathrm{d}\mu = \mathscr{G}\left(\int_{\mathcal{X}} \mathscr{F} \, \mathrm{d}\mu \right).$$

The result follows from Theorem 3.3.14 if we can show that $\mathscr{G} \in \mathfrak{B}(\mathfrak{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2), \mathcal{H}_2)$ since $\mathfrak{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$ is a Hilbert space. This holds since by definition

$$\|\mathscr{G}\| = \sup_{\mathscr{F} \in \mathfrak{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2), \|\mathscr{F}\|_{HS} = 1} \|\mathscr{F}h\|_2 \leqslant \sup_{\mathscr{F} \in \mathfrak{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2), \|\mathscr{F}\|_{HS} = 1} \|\mathscr{F}\|_{HS} \|h\|_1 = \|h\|_1 < \infty,$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denotes the norm in $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively.

□

Probability and statistics on Hilbert spaces

In this chapter we will generalize the usual concepts of probability theory on $\mathbb{R}^d$ to abstract infinite-dimensional Hilbert spaces. We will show how to define random variables with values in infinite-dimensional Hilbert spaces and prove several of their properties. We will also prove the existence of conditional expectations for Hilbertian random variables and their properties. Finally we will introduce simple statistics for Hilbertian random variables with a focus on moment estimators and linear models between Hilbert spaces.

4.1 Hilbertian probability theory

In this section we generalize the well-known ideas from the theory of real-valued random variables to random variables with values in a separable Hilbert space. In all that follows we will only consider separable Hilbert spaces for the reasons mentioned in the previous chapter: countable orthonormal bases and avoiding measurability concerns.

In the usual construction of measure-theoretic probability, random variables are defined as measure functions from a probability space $(\Omega, \mathbb{F}, P)$ into the real numbers with the Borel σ-algebra, $(\mathbb{R}, \mathbb{B})$. This mapping then defines a push-forward probability measure $X(P)$ on $(\mathbb{R}, \mathbb{B})$ such that

$$X(P)(B) = P(X \in B) = P(\{\omega \in \Omega \mid X(\omega) \in B\}) \quad \text{for all } B \in \mathbb{B}.$$

This measure is then referred to as the distribution of the random variable. We will echo this construction by defining random variables as Borel measurable mappings from the probability space $(\Omega, \mathbb{F}, P)$ into $(\mathcal{H}, \mathbb{B}(\mathcal{H}))$.

Obtaining an intuition about a σ-algebra may be difficult but it is often helpful to know some generators of the σ-algebra to discover which sets are "fundamental" to the σ-algebra. By definition $\mathbb{B}(\mathcal{H})$ is generated by the open sets on $\mathcal{H}$ but even that is not particularly helpful,

since these are in themselves quite unwieldy. It turns out that the σ-algebra is also generated by the pre-images of the open sets on $\mathbb{R}$ under the linear functionals on $\mathcal{H}$.

Theorem 4.1.1 (Generator of $\mathbb{B}(\mathcal{H})$)**.** Let $\mathcal{H}$ be a Hilbert space, let $\mathcal{O}$ denote the open sets on $\mathbb{R}$ and for each $h \in \mathcal{H}$ define $\phi_h(x) = \langle x, h \rangle$, i.e. ϕ_h is the linear functional associated with h through Riesz representation theorem. Let $\mathcal{M}$ be the family of sets given by

$$\mathcal{M} = \{\phi_h^{-1}(O) \mid h \in \mathcal{H}, O \in \mathcal{O}\}.$$

Then $\sigma(\mathcal{M}) = \mathbb{B}(\mathcal{H})$.

See [12] for a proof of this. We can now state a handy characterization of measurability wrt. $(\mathcal{H}, \mathbb{B}(\mathcal{H}))$.

Theorem 4.1.2 (Measurability wrt. $\mathbb{B}(\mathcal{H})$ and distributions on $\mathcal{H}$)**.** Let X be a mapping from some probability space $(\Omega, \mathbb{F}, P)$ into the separable Hilbert space $\mathcal{H}$ with the Borel σ-algebra: $(\mathcal{H}, \mathbb{B}(\mathcal{H}))$. Then

1. X is measurable if and only if $\langle X, h \rangle$ is measurable for all $h \in \mathcal{H}$.
2. If X is measurable, its distribution is uniquely determined by the marginal distributions of $\langle X, h \rangle$ for $h \in \mathcal{H}$.

See [12] for a proof. Theorem 4.1.2 states, that we can transform many of our problems on $\mathcal{H}$ to problems on $\mathbb{R}$, where we have a large and well-known toolbox of results to apply. We can now define a Hilbertian random variable:

Definition 4.1.3 (Hilbertian random variable)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $(\mathcal{H}, \mathbb{B}(\mathcal{H}))$ denote the measurable space consisting of a Hilbert space $\mathcal{H}$ and the Borel σ-algebra on $\mathcal{H}$. A measurable mapping $X : \Omega \to \mathcal{H}$ is denoted a *Hilbertian random variable*.

We saw in Theorem 4.1.2 that a Hilbertian random variable X is characterized uniquely by applying the inner product to X and elements of $\mathcal{H}$. By Riesz representation theorem this amounts knowing the distribution of $\phi(X)$ for all $\phi \in \mathcal{H}^*$. Thus the behaviour of the linear functionals on X uniquely determines the distribution and this leads to following definition.

Definition 4.1.4 (Gaussian Hilbertian random variables)**.** Let X be a Hilbertian random variable on $\mathcal{H}$. We say that X is *Gaussian* or *normal* if $\langle X, h \rangle$ is normally distributed (in the usual sense) for all $h \in \mathcal{H}$.

We are used to characterizing a Gaussian random variable by its mean and variance on $\mathbb{R}$, so ideally we would like to find something akin to a mean and a variance for Hilbertian random

variables to characterize Gaussians on $\mathcal{H}$. These should be well-behaved when applying the linear functionals, so that we can easily find the mean and variance of $\langle X, h\rangle$ for each $h \in \mathcal{H}$.

Let us first consider the mean. We would like to construct a functional of functionals – a functional that takes an element $\langle \cdot, h\rangle$ of $\mathcal{H}^*$ and returns the mean of $\langle X, h\rangle$. It is easy to see that from the linearity of the usual expectation on $\mathbb{R}$ that this is a linear functional on $\mathcal{H}^*$. Note also that by the triangle inequality for integrals and Cauchy-Schwarz inequality, we have

$$|E(\langle X, h\rangle)| \leqslant E|\langle X, h\rangle| \leqslant E\|X\|\|h\|.$$

This implies that the functional is bounded if the norm of X has finite first moment. We could instead view this as a mapping from $\mathcal{H}$ into $\mathbb{R}$ that sends h to $E(\langle X, h\rangle)$ which would still be a bounded linear functional, since the inner product is bilinear, thus the mean is also a bounded linear functional on $\mathcal{H}$. By Riesz representation theorem this implies that there exists a unique representer $\mu \in \mathcal{H}$ so that we can express the mean of $\langle X, h\rangle$ simply as $\langle \mu, h\rangle$ for any $h \in \mathcal{H}$.

This is an implicit definition of the mean (which would define the Pettis integral of X) but we have already developed the theory of Bochner integration, so we instead define the mean of X as a "weighted" average of the outcomes of the random variable, just as it was done in the univariate case. This leads to the following definition.

Definition 4.1.5 (Mean of Hilbertian random variable)**.** Let X be a Hilbertian random variable and assume that $E\|X\| < \infty$. The *mean element* or *expectation* of X is given by the Bochner integral

$$E(X) := \int_\Omega X \, \mathrm{d}P.$$

It turns out that the implicit definition and the Bochner definition are the same in our case.

Theorem 4.1.6 (Characterization of the mean of Hilbertian random variable)**.** Let X be a Hilbertian random variable with values in $\mathcal{H}$ and assume that $E\|X\| < \infty$. Let $\mu = E(X)$. Then for any $h \in \mathcal{H}$

$$\langle \mu, h\rangle = E(\langle X, h\rangle).$$

Proof.
The proof is straightforward by noting that the inner product defines a linear functional for each $h \in \mathcal{H}$ and then the result follows from Theorem 3.3.14. □

The definition of the mean using the Bochner integral is preferred over the implicit definition, since we have theorems stating the behaviour of the Bochner integral when composing with linear functionals as applied in the previous proof. The derivation above leads to the nice

dual view of the mean as both a measure of central tendency and as the representer for the functional that takes each h to the mean of $\langle X, h\rangle$.

We are left with the issue of computing the variance of $\langle X, h\rangle$ for any $h \in \mathcal{H}$. We could try to define a functional as with the implicit definition of the mean that maps h to $\mathrm{Var}(\langle X, h\rangle)$ but unfortunately this is not a linear operation. Note however that this is a quadratic form, since if we define the bilinear form $(h, k) \mapsto \mathrm{Cov}(\langle X, h\rangle, \langle X, k\rangle)$ then the variance is simply the bilinear form applied to (h, h). The bilinear form is nicer than the quadratic form but note that if the bilinear form is bounded, there exists an operator $\mathscr{K} \in \mathfrak{B}(\mathcal{H})$ so that the bilinear form can be written $\langle \mathscr{K} h, k\rangle$. A calculation similar to the one done for the mean will show that the bilinear form is bounded if $\|X\|$ has finite second moment. We can deduce the exact form of the aforementioned operator, since by properties of the inner product and the just proven property of the mean of a Hilbertian random variable, we get

$$\begin{aligned}\mathrm{Cov}(\langle X, h\rangle, \langle X, k\rangle) &= E\left[(\langle X, h\rangle - \langle \mu, h\rangle)(\langle X, k\rangle - \langle \mu, k\rangle)\right] = E(\langle X - \mu, h\rangle\langle X - \mu, k\rangle) \\ &= E(\langle\langle X - \mu, h\rangle(X - \mu), k\rangle) = \langle E(\langle X - \mu, h\rangle(X - \mu)), k\rangle.\end{aligned}$$

Thus the operator is defined by the relation $\mathscr{K} h = \langle E(\langle X - \mu, h\rangle(X - \mu))$. Note that since $(X - \mu) \odot (X - \mu) = \langle X - \mu, \cdot\rangle(X - \mu)$, we can calculate the mean above using a Bochner integral over $\mathfrak{B}_{HS}(\mathcal{H})$ which we have shown is a separable Hilbert space, whenever $\mathcal{H}$ is separable. This leads to the following definition.

Definition 4.1.7 (Covariance operator of Hilbertian random variable)**.** Let X be a Hilbertian random variable and assume $E\|X\|^2 < \infty$. Let $\mu = E(X)$. We define the *covariance operator* of X as the Bochner integral

$$\mathrm{Cov}(X) := E((X - \mu) \odot (X - \mu)) = \int_\Omega (X - \mu) \odot (X - \mu)\, \mathrm{d}P.$$

It is not immediately obvious how the outer product and the expectation interact so to add some intuition, we prove the following theorem.

Theorem 4.1.8 (Expectation of outer product of independent variables)**.** Let X_1 and X_2 be Hilbertian random variables on $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively.

Then for any $h \in \mathcal{H}_1$

$$E(X_1 \odot_1 X_2)h = E((X_1 \odot_1 X_2)h),$$

and if $X_1 \perp\!\!\!\perp X_2$, we have

$$E(X_1 \odot_1 X_2) = E(X_1) \odot_1 E(X_2).$$

Proof.
The first claim follows immediately from Theorem 3.3.15. Letting $\langle \cdot, \cdot \rangle$ denote the inner

product on $\mathcal{H}_1$, the second claim holds since for every $h \in \mathcal{H}_1$

$$\begin{aligned} E(X_1 \odot_1 X_2)h = E((X_1 \odot_1 X_2)h) &= E(\langle h, X_1\rangle X_2) \\ &= E(\langle X_1, h_1\rangle)E(X_2) = (E(X_1) \odot_1 E(X_2))h, \end{aligned}$$

where the second to last inequality is due to the independence of X_1 and X_2. □

Note that Theorem 4.1.8 still holds when X_2 is not random. Let us now prove that the covariance does in fact satisfy the implicit definition given in the motivation and some other properties of the covariance operator.

Theorem 4.1.9 (Properties of covariance operator)**.** Let X be a Hilbertian random variable and assume $E\|X\|^2 < \infty$. Let $\mu = E(X)$ and $\mathscr{K}$ be the covariance operator of X. Then

1. $\langle \mathscr{K} h, k\rangle = \mathrm{Cov}(\langle X, h\rangle, \langle X, k\rangle)$.
2. $\mathscr{K}$ is non-negative-definite and trace-class with $\|\mathscr{K}\|_{TR} = \mathrm{Var}\|X\|$.
3. $\mathscr{K} = E(X \odot X) - \mu \odot \mu$.
4. $\mathscr{K} = 0$ if and only if $P(X = \mu) = 1$.

Proof.

1. By using Theorem 4.1.8, Theorem 4.1.6 and various properties of the inner product, we get

$$\begin{aligned} \langle \mathscr{K} h, k\rangle = \langle E[\langle h, X - \mu\rangle(X - \mu)], k\rangle &= E[\langle h, X - \mu\rangle\langle (X - \mu), k\rangle] \\ &= E[(\langle X, h\rangle - \langle \mu, h\rangle)(\langle X, k\rangle - \langle \mu, k\rangle)] \\ &= E[(\langle X, h\rangle - E[\langle X, h\rangle])(\langle X, k\rangle - E[\langle X, k\rangle])] \\ &= \mathrm{Cov}(\langle X, h\rangle, \langle X, k\rangle), \end{aligned}$$

   which proves the result.

2. From the previous claim, we can see that the covariance is non-negative-definite. Letting $(e_n)_{n\in\mathbb{N}}$, we also use the previous claim to calculate the trace norm and get

$$\|\mathscr{K}\|_{TR} = \sum_{i=1}^{\infty} \langle \mathscr{K} e_i, e_i\rangle = \sum_{i=1}^{\infty} \mathrm{Var}(\langle X, e_i\rangle).$$

   The sum of variances is dominated by the sum of second moments of $\langle X, e_i\rangle$ which by Parseval's identity is exactly $E\|X\|^2$ which we have assumed to be finite. This shows

that the sum is finite and algebraic manipulations are sensible. To get the exact value we will juggle expectations and summations using Theorem 3.3.13 and apply Parseval's identity to get

$$\begin{aligned}\|\mathscr{K}\|_{TR} &= \sum_{i=1}^{\infty} \mathrm{Var}(\langle X, e_i\rangle) = \sum_{i=1}^{\infty} E\left[(\langle X, e_i\rangle - \langle \mu, e_i\rangle)^2\right] \\ &= E\left[\sum_{i=1}^{\infty}\langle X, e_i\rangle^2\right] + \sum_{i=1}^{\infty}\langle \mu, e_i\rangle^2 - 2E\left[\sum_{i=1}^{\infty}\langle \mu, e_i\rangle\langle X, e_i\rangle\right] \\ &= E\|X\|^2 - \|\mu\|^2 = \mathrm{Var}(\|X\|),\end{aligned}$$

as desired.

3. By linearity of the outer product and expectation, we have

$$\mathscr{K} = E(X \odot X) - E(X \odot \mu) - E(\mu \odot X) + E(\mu \odot \mu).$$

The final term is not random, so is simply equal to $\mu \odot \mu$. Note that for $h \in \mathcal{H}$, we have

$$E(X \odot \mu)h = E(\langle h, X\rangle \mu) = \langle \mu, h\rangle \mu = (\mu \odot \mu)h,$$

and similarly for the term $E(\mu \odot X)$, so both are equal to $\mu \odot \mu$, which proves the statement.

4. If $X = \mu$ a.s. we can partition the integral in the definition of covariance operator into a region where $X = \mu$ and one where $X \neq \mu$ and get that $\mathscr{K}$ is zero.

Assume instead that $\mathscr{K}$ is 0. Then by the first claim for any $h \in \mathcal{H}$,

$$0 = \langle \mathscr{K} h, h\rangle = \mathrm{Var}(\langle X, h\rangle),$$

which implies that $\langle X, h\rangle$ is equal to $E(\langle X, h\rangle) = \langle \mu, h\rangle$ almost surely. Let $(e_i)_{i\in\mathbb{N}}$ be an ONB for $\mathcal{H}$ and set A equal to the set of $\omega \in \Omega$ where $\langle X(\omega), e_i\rangle = \langle \mu, e_i\rangle$ for all $i \in \mathbb{N}$. Countable intersections of almost sure sets are almost sure so $P(A) = 1$. Then for each $\omega \in A$, we have by applying the Fourier expansion

$$\mu = \sum_{i=1}^{\infty}\langle \mu, e_i\rangle e_i = \sum_{i=1}^{\infty}\langle X(\omega), e_i\rangle e_i = X(\omega),$$

so $X = \mu$ almost surely.

□

With these definitions in mind we can note that just like in the univariate and multivariate cases, we can characterize Gaussian distributions by their mean and covariance operator.

Theorem 4.1.10 (Characterization of Hilbertian Gaussian random variables)**.** Let $\mathcal{H}$ be a Hilbert space. Given any $\mu \in \mathcal{H}$ and non-negative-definite $\mathscr{K} \in \mathfrak{B}_{TR}(\mathcal{H})$ there exists a Hilbertian random variable X that is Gaussian with mean μ and covariance $\mathscr{K}$. Conversely, if X is Gaussian, X has finite second moment so that the mean and covariance of X exist. We denote this Gaussian with $\mathcal{N}(\mu, \mathscr{K})$.

Proof.
A proof can be provided by using characteristic functionals, see [28] Proposition 2.7, 2.8 and Theorem IV.2.4. □

Remark 4.1.11 (Non-existence of infinite-dimensional standard Gaussian)**.** If we assume that $\mathcal{H}$ is infinite-dimensional, then we saw earlier that the identity operator $\mathscr{I}$ is not compact and hence not a trace-class operator. Therefore there does not exist a Gaussian Hilbertian random variable with covariance operator equal to the identity operator. We're used to the idea of being given some arbitrary normal distribution and then "whitening" it by subtracting the mean and multiplying by the square root of the inverse of the covariance to get a standard normal distribution. This procedure is no longer available to us (since also the covariance is non-invertible) and as such no infinite-dimensional Gaussian is the "reference" Gaussian as is the case in finite dimensions.

We can also note that just as for independent finite-dimensional Gaussian variables, we can form linear combinations of independent infinite-dimensional Gaussian variables and retain the Gaussian distribution.

Theorem 4.1.12 (Linear combinations of independent Gaussians are Gaussian)**.** Let X and Y be independent Gaussian random variables on Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively. Let μ_X and μ_Y denote the mean of X and Y respectively and let $\mathscr{K}_X$ and $\mathscr{K}_Y$ denote the covariance operators. Let $\mathscr{A}_X$ and $\mathscr{A}_Y$ be bounded operators to a third Hilbert space $\mathcal{H}$ from $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively.

Then $\mathscr{A}_X X + \mathscr{A}_Y Y$ is a Hilbertian Gaussian with mean $\mu = \mathscr{A}_X \mu_X + \mathscr{A}_Y \mu_Y$ and covariance operator $\mathscr{K} = \mathscr{A}_X \mathscr{K}_X \mathscr{A}_X^* + \mathscr{A}_Y \mathscr{K}_Y \mathscr{A}_Y^*$.

Proof.
For a proof see [18] Proposition 4.8 and 4.9. □

Earlier we showed how to construct an eigen-decomposition for non-negative operators, which we can now use to decompose the covariance operator.

Theorem 4.1.13 (Eigen-decomposition of covariance operator)**.** Let X be a Hilbertian random variable with second moment and let $\mathscr{K}$ denote the covariance of X. Then $\mathscr{K}$ admits an eigen-decomposition

$$\mathscr{K} = \sum_{j=1}^{\infty} \lambda_j e_j \odot e_j,$$

where $(e_j)_{j=1}^{\infty}$ is an orthonormal basis for $\overline{\mathrm{Im}(\mathscr{K})}$ and the eigenvalues $(\lambda_j)_{j=1}^{\infty}$ are non-negative and tending to zero with each eigenvalue having finite multiplicity.

Using the orthonormal basis from the decomposition of the covariance operator lets us decompose a Hilbertian random variable into a sequence of real-valued random variables as seen below.

Theorem 4.1.14 (Fourier expansion of random variables)**.** Let X be a Hilbertian random variable with second moment and let μ and $\mathscr{K}$ denote the mean and covariance of X respectively. Let further $(\lambda_j)_{j=1}^{\infty}$ and $(e_j)_{j=1}^{\infty}$ be the eigenvalues and -vectors of the covariance operator. Then with probability 1, we have

$$X = \sum_{j=1}^{\infty} \langle X, e_j \rangle e_j,$$

where $(\langle X, e_j \rangle)_{j=1}^{\infty}$ are uncorrelated real-valued random variables with mean $\langle \mu, e_j \rangle$ and variances λ_j.

Note that the above is an extension of the principal components decomposition of multivariate random variables to the context of Hilbertian random variables. Those familiar with the theory of stochastic processes will also note the similarity to the Karhunen-Loève decomposition (as given in [12] Theorem 7.3.5). Using this decomposition we can derive the distribution of the norm of a mean-zero Gaussian random variable, which we will use later.

Theorem 4.1.15 (Distribution of the norm of Hilbertian Gaussian)**.** Let X be a Hilbertian random variable with $X \sim \mathcal{N}(\mu, \mathscr{K})$ on the space $\mathcal{H}$ with norm $\|\cdot\|$. Then letting $(\lambda_j)_{j=1}^{\infty}$ be the eigenvalues of $\mathscr{K}$ and $(Z_n)_{n \in \mathbb{N}}$ be a sequence of iid. standard normal random variables, we get

$$\|X - \mu\|^2 \overset{\mathcal{D}}{=} \sum_{i=1}^{\infty} \lambda_i Z_i^2.$$

Proof.
Using the previous theorem, we can use the eigen-decomposition of $\mathscr{K}$ and write

$$X - \mu = \sum_{j=1}^{\infty} \langle X - \mu, e_j \rangle e_j.$$

Thus taking norms on either side and recalling Parseval's identity, we get using simple properties of the norm and the fact that $\langle X - \mu, e_j \rangle \sim \mathcal{N}(0, \lambda_j)$

$$\|X - \mu\|^2 = \left\| \sum_{j=1}^{\infty} \langle X - \mu, e_j \rangle e_j \right\|^2 = \sum_{j=1}^{\infty} \|\langle X - \mu, e_j \rangle e_j\|^2 \overset{\mathcal{D}}{=} \sum_{j=1}^{\infty} \lambda_i Z_i^2,$$

as desired. □

This will be applied later when constructing test statistics.

Having defined the mean and covariance of a single Hilbertian random variable, it is natural when we need to consider questions of independence and conditional independence to ask how two Hilbertian random variables behave together. Letting X_1 and X_2 denote two Hilbertian random variables with values in separable Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, we can once again consider the implications of Theorem 4.1.2 and settle for characterizing how $\langle X_1, h_1 \rangle_1$ and $\langle X_2, h_2 \rangle_2$ behave for every $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$. Given two univariate real-valued random variables, we often settle for calculating the covariance of the variables as a measure of correlation. We could now repeat many of the arguments given above the definition of the covariance operator to get a bilinear form, that takes a functional on $\mathcal{H}_1$ and a functional on $\mathcal{H}_2$ and returns the covariance of the functionals applied to the respective random variables. These arguments would lead us to define a cross-covariance operator as below.

Definition 4.1.16 (Cross-covariance operator)**.** Let X_1 and X_2 be Hilbertian random variables on $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively. Assume that both X_1 and X_2 have finite second moment and let μ_1 and μ_2 be the means of X_1 and X_2 respectively. We define the *cross-covariance operator* of X_1 and X_2 as the Bochner integral

$$\operatorname{Cov}(X, Y) := E((X_2 - \mu_2) \odot_2 (X_1 - \mu_1)) = \int_\Omega (X_2 - \mu_2) \odot_2 (X_1 - \mu_1) \, \mathrm{d}P.$$

Note that the integral above is well-defined since the outer product is an element of $\mathfrak{B}_{HS}(\mathcal{H}_2, \mathcal{H}_1)$ which is a separable Hilbert space. Let us prove some properties of the cross-covariance operator.

Theorem 4.1.17 (Properties of cross-covariance operators)**.** Let X_1 and X_2 be Hilbertian random variables on $\mathcal{H}_1$ and $\mathcal{H}_2$ with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ respectively. Assume that both X_1 and X_2 have finite second moment and let μ_1 and μ_2 be the means, $\mathscr{K}_1$ and $\mathscr{K}_2$ be the covariances and $\mathscr{K}_{12}$ the cross-covariance of X_1 and X_2 respectively, i.e. $\mathscr{K}_{12} = \operatorname{Cov}(X_1, X_2)$. Let further $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$. Then

1. $\langle \mathscr{K}_{12} h_2, h_1 \rangle_1 = \operatorname{Cov}(\langle X_1, h_1 \rangle_1, \langle X_2, h_2 \rangle_2)$.
2. $|\langle \mathscr{K}_{12} h_2, h_1 \rangle_1| \leqslant \sqrt{\langle \mathscr{K}_1 h_1, h_1 \rangle} \sqrt{\langle \mathscr{K}_2 h_2, h_2 \rangle}$.

3. $\mathscr{K}_{12}^* = \mathscr{K}_{21} = E((X_1 - \mu_1) \odot_1 (X_2 - \mu_2))$.

4. $\mathscr{K}_{12} = E(X_2 \odot_2 X_1) - \mu_2 \odot_2 \mu_1$.

5. If $X_1 \perp\!\!\!\perp X_2$ then $\mathscr{K}_{12} = 0$.

Proof.
The first four claims follow from arguments similar to Theorem 4.1.9. The final claim can be seen to hold by applying Theorem 4.1.8. □

As with real-valued random variables, we will be interested in sequences of Hilbertian random variables and in particular their convergence properties. To that end let us define the modes of convergence for Hilbertian random variables.

Definition 4.1.18 (Modes of convergence for Hilbertian random variables)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of Hilbertian random variables on $\mathcal{H}$ and let X be another Hilbertian random variable. Let also $\|\cdot\|$ denote the norm on $\mathcal{H}$. Then

1. If $P(\lim_{n\to\infty} X_n = X) = 1$, we say that X_n *converges to* X *almost surely* and write $X_n \overset{a.s.}{\to} X$.

2. If for any $\varepsilon > 0$, we have $\lim_{n\to\infty} P(\|X_n - X\| \geqslant \varepsilon) = 0$, we say that X_n *converges to* X *in probability* and write $X_n \overset{P}{\to} X$.

3. If for any continuous, bounded function $f : \mathcal{H} \to \mathbb{R}$,

$$E(f(X_n)) \to E(f(X)), \quad \text{as } n \to \infty,$$

we say that X_n *converges in distribution to* X and write $X_n \overset{\mathcal{D}}{\to} X$.

These are straight-forward generalizations of the usual definitions for real-valued random variables. We will omit a full disposition of these modes of convergence for Hilbertian random variables and simply note, that we have almost all the results we're used to for real random variables. In particular both the continuous mapping theorem and Slutsky's theorem still hold (for proofs see for instance Theorem 2.7 and 3.1 in [2]).

Theorem 4.1.19 (Continuous mapping theorem)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of Hilbertian random variables on $\mathcal{H}$ and let X be another Hilbertian random variable. Assume that $X_n \overset{\mathcal{D}}{\to} X$. Assume further $\tilde{\mathcal{H}}$ is another Hilbert space and that $f : \mathcal{H} \to \tilde{\mathcal{H}}$ is a continuous mapping. Then $f(X_n) \overset{\mathcal{D}}{\to} f(X)$.

Theorem 4.1.20 (Slutsky's theorem)**.** Let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of Hilbertian random variables with values in $\mathcal{H}$ and let X be another Hilbertian random variable on the space and $h \in \mathcal{H}$. Assume that $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P} h$. Then

$$X_n + Y_n \xrightarrow{\mathcal{D}} X + h.$$

We end the section with two generalizations of the classical large sample results: the Law of Large Numbers and the Central Limit Theorem as seen in [12].

Theorem 4.1.21 (Law of Large Numbers in Hilbert spaces)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of Hilbertian random variables. Assume that the sequence is independent and identically distributed and assume further that the common distribution has finite first moment with mean μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu.$$

Theorem 4.1.22 (Central Limit Theorem in Hilbert spaces)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of Hilbertian random variables. Assume that the sequence is independent and identically distributed and assume further that the common distribution has mean zero and finite second moment. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \xrightarrow{\mathcal{D}} G,$$

where G is a Gaussian Hilbertian random variable with mean zero and covariance operator $\mathrm{Cov}(X) = E(X \odot X)$.

4.2 Conditional expectation for Hilbertian random variables

In this section we develop and define the theory of conditional expectations for Hilbertian random variables.

There are several equivalent ways of defining and proving the existence of conditional expectations for integrable random variables, real-valued or Hilbertian. Most approaches use the Radon-Nikodym theorem to prove existence of the conditional expectation, while others will apply Hilbert space projection techniques. We will apply a more direct approach employing the theory of Bochner integration directly and utilizing that existence of real-valued conditional expectations is already established.

Let us first define the conditional expectation for a Hilbertian random variable.

Definition 4.2.1 (Conditional expectation of Hilbertian random variable)**.** Let X be a Hilbertian random variable defined on $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}$. Let $\mathbb{D} \subseteq \mathbb{F}$ be another σ-algebra. Assume further that X has finite first moment. A *conditional expectation of X given* $\mathbb{D}$ is a Hilbertian random variable, Y, satisfying

1. Y is $\mathbb{D}$-measurable and Bochner integrable.
2. For any $D \in \mathbb{D}$:
$$\int_D X \, \mathrm{d}P = \int_D Y \, \mathrm{d}P.$$

In the following proof we will employ that the set of equivalence classes of Bochner integrable functions on a measure space form a Banach space.

Theorem 4.2.2 (L^1 space for Bochner integrals)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space, $\mathcal{H}$ a Hilbert space and denote by $\mathcal{L}^1(\mathcal{X}, \mathbb{E}, \mu; \mathcal{H})$ the set of all Bochner integrable functions from $\mathcal{X}$ to $\mathcal{H}$ with the Borel σ-algebra. Define $L^1(\mathcal{X}, \mathbb{E}, \mu; \mathcal{H})$ as the set of equivalence classes of $\mathcal{L}^1(\mathcal{X}, \mathbb{E}, \mu; \mathcal{H})$ where two functions are in the same equivalence class, if they are equal except possibly on a null set.

$L^1(\mathcal{X}, \mathbb{E}, \mu; \mathcal{H})$ forms a Banach space, i.e. it is a vector space under pointwise addition and scalar multiplication that is complete with respect to the norm

$$\|f\|_{L^1} = \int_X \|f\|_{\mathcal{H}} \, \mathrm{d}\mu$$

for $f \in L^1(\mathcal{X}, \mathbb{E}, \mu; \mathcal{H})$.

For a proof see [11] Theorem 3.7.7. We now prove that the conditional expectation of a Hilbertian random variable exists and is unique.

Theorem 4.2.3. The conditional expectation of a Hilbertian random variable as defined in Definition 4.2.1 exists and is almost surely unique. We can therefore refer to *the* conditional expectation and denote the conditional expectation of X given $\mathbb{D}$ by $E(X \mid \mathbb{D})$.

Proof.
We follow the proof given in [22] Theorem 2.1. Throughout we denote the norm on $\mathcal{H}$ by $\|\cdot\|_{\mathcal{H}}$ and the norm on $L^1(\Omega, \mathbb{F}, P; \mathcal{H})$ by $\|\cdot\|_{L^1}$.

We have assumed that X has finite first moment, i.e. it is integrable and thus by definition there exists a sequence of simple functions $(X_n)_{n \in \mathbb{N}}$ converging pointwise to X and further

$$\lim_{n \to \infty} \int_\Omega \|X_n - X\|_{\mathcal{H}} \, \mathrm{d}P = 0.$$

This amounts to X_n converging to X in L^1 norm as given in Theorem 4.2.2.

Each X_n can be written

$$X_n(\omega) = \sum_{i=1}^{k} 1_{A_i}(\omega) h_i,$$

for $A_1, \dots A_k \in \mathbb{F}$ and $h_1, \dots h_k \in \mathcal{H}$. We first show that

$$E(X_n \mid \mathbb{D}) = \sum_{i=1}^{k} E(1_{A_i} \mid \mathbb{D}) h_i,$$

where each of the $E(1_{A_i} \mid \mathbb{D})$ are real-valued conditional expectations and thus well-defined. The proposed expression is obviously $\mathbb{D}$-measurable, so we show the integral property and get for $D \in \mathbb{D}$ by simple properties of the Bochner integral and the definition of $E(1_{A_i} \mid \mathbb{D})$ that

$$\int_D E(X_n \mid \mathbb{D}) \, \mathrm{d}P = \sum_{i=1}^{k} \int_D E(1_{A_i} \mid \mathbb{D}) \, \mathrm{d}P h_i = \sum_{i=1}^{k} \int_D 1_{A_i} \, \mathrm{d}P h_i = \int X_n \, \mathrm{d}P.$$

Similar calculations will show that if X_n and X_m are simple functions, we have

$$E(X_n - X_m \mid \mathbb{D}) = E(X_n \mid \mathbb{D}) - E(X_m \mid \mathbb{D}).$$

We also have a triangle inequality, since

$$\|E(X_n \mid \mathbb{D})\|_{\mathcal{H}} = \|\sum_{i=1}^{k} E(1_{A_i} \mid \mathbb{D}) h_i\|_{\mathcal{H}} \leqslant \sum_{i=1}^{k} E(1_{A_i} \mid \mathbb{D}) \|h_i\|_{\mathcal{H}} = E(\|X_n\| \mid \mathbb{D}).$$

Using the results above, we get

$$\begin{aligned} \|E(X_n \mid \mathbb{D}) - E(X_m \mid \mathbb{D})\|_{L^1} &= \|E(X_n - X_m \mid \mathbb{D})\|_{L^1} \leqslant E\left[E(\|X_n - X_m\|_{\mathcal{H}} \mid \mathbb{D})\right] \\ &= E(\|X_n - X_m\|_{\mathcal{H}}) = \|X_n - X_m\|_{L^1} \end{aligned}$$

where the last term goes to zero as $n, m \to \infty$ by construction, thus $E(X_n \mid \mathbb{D})$ is a Cauchy sequence in $L^1(\Omega, \mathbb{D}, P; \mathcal{H})$. The space is complete by Theorem 4.2.2 so $E(X_n \mid \mathbb{D})$ converges and we define the limit to be the conditional expectation of X given $\mathbb{D}$, i.e. $E(X \mid \mathbb{D}) := \lim_{n \to \infty} E(X_n \mid \mathbb{D})$. The limit is almost surely unique and measurable with respect to $\mathbb{D}$ by construction, so we just need to show the integral property.

For this note that for $D \in \mathbb{D}$, we have by the triangle inequality

$$\begin{aligned} &\left\| \int_D X \, \mathrm{d}P - \int_D E(X \mid \mathbb{D}) \, \mathrm{d}P \right\|_{\mathcal{H}} \\ &\leqslant \left\| \int_D X \, \mathrm{d}P - \int_D E(X_n \mid \mathbb{D}) \, \mathrm{d}P \right\|_{\mathcal{H}} + \left\| \int_D E(X_n \mid \mathbb{D}) \, \mathrm{d}P - \int_D E(X \mid \mathbb{D}) \, \mathrm{d}P \right\|_{\mathcal{H}}. \end{aligned}$$

Note that the second integral in the first term is equal to $\int_D X_n \, \mathrm{d}P$, since each X_n is simple and we showed already that conditional expectations exist for simple functions. Therefore the first term becomes the L^1 norm of $X - X_n$ which goes to 0 by construction. The second term goes to zero by definition of $E(X \mid \mathbb{D})$ – the term is exactly the L^1 norm of $E(X \mid \mathbb{D}) - E(X_n \mid \mathbb{D})$. □

The definition of a conditional expectation is identical to the one given for real-valued random variables and thus it is not surprising that many of the same properties apply. We get most of the properties except the ones given by multiplication or monotonicity since unlike $\mathbb{R}$, $\mathcal{H}$ has no multiplication and is not ordered.

Theorem 4.2.4 (Properties of Hilbertian conditional expectation)**.** Let X be a Hilbertian random variable defined on $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}$. Let $\mathbb{D} \subseteq \mathbb{F}$ be another σ-algebra and assume the first moment of X is finite. Then

1. $E\|E(X \mid \mathbb{D})\| \leqslant E\|X\|$
2. If $\mathbb{D} \subseteq \mathbb{E}$ are sub-σ-algebras of $\mathbb{F}$, we have $E(X \mid \mathbb{D}) = E(E(X \mid \mathbb{D}) \mid \mathbb{E}) = E(E(X \mid \mathbb{E}) \mid \mathbb{D})$.
3. If X is $\mathbb{D}$-measurable, then $E(X \mid \mathbb{D}) = X$.
4. If X is independent of $\mathbb{D}$, then $E(X \mid \mathbb{D}) = E(X)$.
5. If $\tilde{\mathcal{H}}$ is another Hilbert space and $\mathscr{A} \in \mathfrak{B}(\mathcal{H}, \tilde{\mathcal{H}})$ then $E(\mathscr{A}X \mid \mathbb{D}) = \mathscr{A}(E(X \mid \mathbb{D}))$.
6. For any $h \in \mathcal{H}$, $E(\langle X, h\rangle \mid \mathbb{D}) = \langle E(X \mid \mathbb{D}), h\rangle$
7. If $\tilde{\mathcal{H}}$ is another Hilbert space and $\mathscr{A}$ is random variable with values in $\mathfrak{B}(\mathcal{H}, \tilde{\mathcal{H}})$ then for any $h \in \mathcal{H}$, $E(\mathscr{A}h \mid \mathbb{D}) = E(\mathscr{A} \mid \mathbb{D})h$.
8. If $(X_n)_{n\in\mathbb{N}}$ is a sequence of integrable Hilbertian random variables, then $E\left(\sum_{n=1}^{\infty} X_n \mid \mathbb{D}\right) = \sum_{n=1}^{\infty} E\left(X_n \mid \mathbb{D}\right)$.

Proof.
Most of these proofs proceed in the same manner. We will illustrate the idea by proving 5. We need to show that $\mathscr{A}(E(X \mid \mathbb{D}))$ satisfies the requirements of being the conditional expectation of $\mathscr{A}X$ given $\mathbb{D}$. Measurability follows trivially, since $\mathscr{A}$ is continuous and pre-images under continuous mappings of Borel sets are again Borel. Integrability follows by

$$E(\|\mathscr{A}(E(X \mid \mathbb{D}))\|_{\tilde{\mathcal{H}}}) \leqslant E(\|\mathscr{A}\|\|E(X \mid \mathbb{D})\|_{\mathcal{H}})$$

since $\mathscr{A}$ is bounded and $E(X \mid \mathbb{D})$ is integrable. The conditional expectation and $\mathscr{A}X$ agree on sets in $\mathbb{D}$ by Theorem 3.3.14 since

$$\int_D \mathscr{A}X \,\mathrm{d}P = \mathscr{A}\int_D X \,\mathrm{d}P = \mathscr{A}\int_D E(X \mid \mathbb{D}) \,\mathrm{d}P = \int_D \mathscr{A}(E(X \mid \mathbb{D})) \,\mathrm{d}P$$

thus proving the result.

The remaining results go through the same steps; arguing for $\mathbb{D}$-measurability using well-known measurability arguments and then applying properties of the Bochner integral to show th integral property. □

Note that the first result proves that the conditional expectation is a contraction, i.e. taking a conditional expectation of a random variable always reduces the norm. This further implies that if the original variable is integrable, so is the conditional expectation, thus we can omit showing integrability, when proving that a random variable is a conditional expectation.

A useful property of the real-valued conditional expectation is "pulling out what is known", i.e. the fact that if X is $\mathbb{D}$-measurable and $E|XY| < \infty$, we have $E(XY \mid \mathbb{D}) = XE(Y \mid \mathbb{D})$. As we already noted, there is no multiplication on $\mathcal{H}$, we do however have scalar multiplication and both an inner and an outer product, that do satisfy this "pulling out what is known"-property, as we shall see below.

Theorem 4.2.5 ("Pulling out what is known")**.** Let X and Y be Hilbertian random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}$, let Z be a Hilbertian random variable on the same probability space but with values in $\tilde{\mathcal{H}}$ and let W be a real-valued random variable on the same probability space. Assume that all the aforementioned random variables have first moment. Let $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$ and denote the inner product of $\mathcal{H}$ by $\langle \cdot, \cdot \rangle$ and the norms of $\mathcal{H}$ and $\tilde{\mathcal{H}}$ by $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\tilde{\mathcal{H}}}$ respectively. Then

1. $E(W \cdot X \mid \mathbb{D}) = W \cdot E(X \mid \mathbb{D})$ if W is $\mathbb{D}$-measurable and $E\|W \cdot X\|_{\mathcal{H}} < \infty$.
2. $E(W \cdot X \mid \mathbb{D}) = E(W \mid \mathbb{D}) \cdot X$ if X is $\mathbb{D}$-measurable and $E\|W \cdot X\|_{\mathcal{H}} < \infty$.
3. $E(\langle X, Y \rangle \mid \mathbb{D}) = \langle X, E(Y \mid \mathbb{D}) \rangle$ if X is $\mathbb{D}$-measurable and $E\left(\|X\|_{\mathcal{H}} \|Y\|_{\mathcal{H}}\right) < \infty$.
4. $E(X \odot_{\mathcal{H}} Z \mid \mathbb{D}) = X \odot_{\mathcal{H}} E(Z \mid \mathbb{D})$ if X is $\mathbb{D}$-measurable and $E\left(\|X\|_{\mathcal{H}} \|Z\|_{\tilde{\mathcal{H}}}\right) < \infty$.
5. $E(Z \odot_{\tilde{\mathcal{H}}} X \mid \mathbb{D}) = E(Z \mid \mathbb{D}) \odot_{\tilde{\mathcal{H}}} X$ if X is $\mathbb{D}$-measurable and $E\left(\|X\|_{\mathcal{H}} \|Z\|_{\tilde{\mathcal{H}}}\right) < \infty$.

Proof.
We proceed to show that the proposed conditional expectations satisfy the requirements given in the definition, i.e. $\mathbb{D}$-measurability and integrals agreeing on all $\mathbb{D}$-sets. Throughout the following let $(e_i)_{i=1}^{\infty}$ be an orthonormal basis for $\mathcal{H}$.

For the first claim note that trivially $W \cdot E(X \,|\mathbb{D})$ is $\mathbb{D}$-measurable. To prove that the integral property holds, note first that for every $D \in \mathbb{D}$

$$\int_D W \cdot X \,\mathrm{d}P = \int_D \sum_{i=1}^{\infty} \langle W \cdot X, e_i \rangle \cdot e_i \,\mathrm{d}P = \sum_{i=1}^{\infty} \int_D W \langle X, e_i \rangle \,\mathrm{d}P \cdot e_i,$$

where we have expanded X using it's Fourier expansion and applied Theorem 3.3.13. Each of the real-valued random variables $W\langle X, e_i\rangle$ has finite first moment by the moment assumption on $W \cdot X$ and Cauchy-Schwarz, thus there exists a real-valued conditional expectation $E(W\langle X, e_i\rangle \mid \mathbb{D})$ whose integrals agree with $W\langle X, e_i\rangle$ on $\mathbb{D}$-sets. Furthermore, W is $\mathbb{D}$-measurable and can thus be pulled outside the integral and we get

$$\begin{aligned}\int_D W \cdot X \,\mathrm{d}P &= \sum_{i=1}^{\infty} \int_D W \langle X, e_i \rangle \,\mathrm{d}P \cdot e_i = \sum_{i=1}^{\infty} \int_D W E(\langle X, e_i \rangle \mid \mathbb{D}) \,\mathrm{d}P \cdot e_i \\ &= \int_D W \cdot \sum_{i=1}^{\infty} \langle E(X \mid \mathbb{D}), e_i \rangle \cdot e_i \,\mathrm{d}P = \int_D W \cdot E(X \mid \mathbb{D}) \,\mathrm{d}P\end{aligned}$$

again by the Fourier expansion, Theorem 3.3.13 and by the way that conditional expectations interact with the inner product. This proves the claim.

The second claim can be proven in a manner analogous to the first claim.

For the third claim note that $\langle X, E(Y \mid \mathbb{D})\rangle$ is trivially measurable if the inner product is measurable, which follows if it is continuous. This can be seen by bilinearity of the inner product, Cauchy-Schwarz and the triangle inequality, since letting $x_n \to x$ and $y_n \to y$ as $n \to \infty$, we get

$$\begin{aligned}&|\langle x_n, y_n \rangle - \langle x, y \rangle| = |\langle x_n, y_n \rangle - \langle x_n, y \rangle + \langle x_n, y \rangle - \langle x, y \rangle| \\ &\leqslant |\langle x_n, y_n - y \rangle| + |\langle x_n - x, y \rangle| \leqslant \|x_n\| \|y_n - y\| + \|x_n - x\| \|y\| \to 0,\end{aligned}$$

proving continuity. To show that the integrals agree on $\mathbb{D}$-sets, we proceed as above and get

$$\int_D \langle X, Y \rangle \,\mathrm{d}P = \sum_{i=1}^{\infty} \int_D \langle X, e_i \rangle \langle Y, e_i \rangle \,\mathrm{d}P$$

by the Fourier expansion of the inner product and Theorem 3.3.13. We can note that the integrand $\langle X, e_i\rangle\langle Y, e_i\rangle$ has finite first moment by Cauchy-Schwarz and the assumption of first moment of $\|X\|_{\mathcal{H}}\|Y\|_{\mathcal{H}}$. Thus there exists a conditional expectation $E(\langle X, e_i\rangle\langle Y, e_i\rangle \mid \mathbb{D})$ that agrees with $\langle X, e_i\rangle\langle Y, e_i\rangle$ on $\mathbb{D}$-sets. Furthermore, we can pull out $\langle X, e_i\rangle$ since X is assumed $\mathbb{D}$-measurable, so any measurable function of X is also $\mathbb{D}$-measurable. Therefore we

get

$$\int_D \langle X, Y\rangle \,\mathrm{d}P = \sum_{i=1}^{\infty} \int_D \langle X, e_i\rangle E(\langle Y, e_i\rangle \mid \mathbb{D}) \,\mathrm{d}P$$
$$= \int_D \sum_{i=1}^{\infty} \langle X, e_i\rangle \langle E(Y \mid \mathbb{D}), e_i\rangle \,\mathrm{d}P = \int_D \langle X, E(Y \mid \mathbb{D})\rangle \,\mathrm{d}P$$

again using the Fourier expansion of the inner product, Theorem 3.3.13 and the way conditional expectations and inner products interact. This proves the claim.

For the fourth claim note as above that if the outer product is continuous, then $X \odot_{\mathcal{H}} Z$ is trivially $\mathbb{D}$-measurable. An argument analogous to the one for the previous claim can show that the outer product is continuous, this time using that we can explicitly calculate the Hilbert-Schmidt norm of the outer product as the product of norms of the arguments instead of Cauchy-Schwarz.

We will show that the integrals agree by showing that the resulting operator performs the same operation on all $h \in \mathcal{H}$. For any $D \in \mathbb{D}$ we get by Theorem 3.3.15

$$\left(\int_D X \odot_{\mathcal{H}} Z \,\mathrm{d}P\right) h = \int_D (X \odot_{\mathcal{H}} Z) h \,\mathrm{d}P = \int_D \langle h, X\rangle Z \,\mathrm{d}P.$$

The integrand $\langle h, X\rangle Z$ has finite first moment by Cauchy-Schwarz and the assumption of first moment of $\|X\|_{\mathcal{H}}\|Z\|_{\tilde{\mathcal{H}}}$. Thus there exists a conditional expectation $E(\langle h, X\rangle Z \mid \mathbb{D})$ that agrees with $\langle h, X\rangle Z$ on $\mathbb{D}$-sets. Furthermore, we can pull out $\langle h, X\rangle$ by the first claim, since X is $\mathbb{D}$-measurable. Thus we get

$$\int_D \langle h, X\rangle Z \,\mathrm{d}P = \int_D \langle h, X\rangle E(Z \mid \mathbb{D}) \,\mathrm{d}P = \left(\int_D X \odot_{\mathcal{H}} E(Z \mid \mathbb{D}) \,\mathrm{d}P\right) h$$

by Theorem 3.3.15 as desired.

The fifth claim can be proven in a manner analogous to the fourth claim. □

We will almost solely be interested in conditional expectations with respect to other random variables, which we will define as below. Note that we do not require the other random variable to be real-valued or even Hilbertian.

Definition 4.2.6 (Hilbertian conditional expectation given random variable)**.** Let X be a Hilbertian random variable defined on $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}$. Let Y be another random variable defined on the same probability space with values in the measure space $(\mathcal{Y}, \mathbb{E})$. Assume that X has first moment. Then we define the *conditional expectation of X given Y* as

$$E(X \mid Y) := E(X \mid \sigma(Y))$$

where $\sigma(Y)$ is the smallest σ-algebra making Y $\mathbb{F}$-$\mathbb{E}$-measurable, which we can write explicitly as

$$\sigma(Y) = \{Y^{-1}(E) \mid E \in \mathbb{E}\}.$$

A Hilbertian random variable X being measurable with respect to the σ-algebra generated by another random variable Y implies that X can be written as a measurable function of Y, as we shall see below.

Theorem 4.2.7 (Doob-Dynkin lemma for Hilbertian random variables)**.** Let X be a Hilbertian random variable defined on $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}$ and let Y be another random variables on the same probability space with values in the measurable space $(\mathcal{Y}, \mathbb{E})$. Then X is $\sigma(Y)$ measurable if and only if there exists a $\mathbb{E} - \mathbb{B}(\mathcal{H})$-measurable function $\phi : \mathcal{Y} \to \mathcal{H}$ so that

$$X = \phi \circ Y.$$

Proof.
Assuming that $X = \phi \circ Y$, it is obvious that X is $\sigma(Y)$ measurable.

For the other implication assume that X is $\sigma(Y)$ measurable and consider the class of Hilbertian random variables given by

$$\mathcal{F} = \{\phi(Y) \mid \phi \text{ is } \mathbb{E} - \mathbb{B}(\mathcal{H})\text{-measurable}\}.$$

If we can show that

1. $Z_1, Z_2 \in \mathcal{F}$ implies that $Z_1 + Z_2 \in \mathcal{F}$,
2. $(Z_n)_{n\in\mathbb{N}} \in \mathcal{F}$ and $Z_n \to Z$ as $n \to \infty$ implies that $Z \in \mathcal{F}$,
3. $1_D \cdot h \in \mathcal{F}$ for all $D \in \sigma(Y)$ and all $h \in \mathcal{H}$,

then we will be done, since any $\sigma(Y)$-measurable random variable can be approximated by a sequence of $\sigma(Y)$-measurable simple random variables.

Assume that $Z_1, Z_2 \in \mathcal{F}$, i.e. $Z_1 = \phi_1(Y)$ and $Z_2 = \phi_2(Y)$, then we can write

$$Z_1 + Z_2 = \phi_1(Y) + \phi_2(Y) = (\phi_1 + \phi_2)(Y),$$

so since sums of measurable mappings are measurable, we have that $Z_1 + Z_2 \in \mathcal{F}$.

Assume now that $(Z_n)_{n\in\mathbb{N}} \in \mathcal{F}$, i.e. each $Z_n = \phi_n(Y)$, and $Z_n \to Z$ as $n \to \infty$. Then we note that $F = (\lim_{n\to\infty} \phi_n \text{ exists})$ is in $\mathbb{E}$, since each ϕ_n is measurable and we can write F using countable intersections and unions. Defining $\phi = \lim_{n\to\infty}(1_F\phi_n)$, we can write

$$Z = \lim_{n\to\infty} Z_n = \lim_{n\to\infty} \phi_n(Y) = \lim_{n\to\infty} (1_F\phi_n)(Y) = \phi(Y),$$

thus proving that $\mathcal{F}$ is closed.

Finally each $Z = 1_D \cdot h$ for some $D \in \sigma(Y)$ and $h \in \mathcal{H}$ is in $\mathcal{F}$, since $D \in \sigma(Y)$ implies that there exists a set $E \in \mathbb{E}$ such that $D = (Y \in E)$ and thus

$$Z = 1_D \cdot h = 1_E(Y) \cdot h,$$

and therefore if $\phi(y) = 1_E(y) \cdot h$ is $\mathbb{E} - \mathbb{B}(\mathcal{H})$-measurable, we will be done. This is obvious since the pre-image of any set $E \in \mathbb{E}$ under ϕ will either be $\{h\}$ or the empty set, both of which are elements of the Borel σ-algebra. □

For conditional expectations this leads to the following definition.

Definition 4.2.8 (Conditional expectation given value of variable)**.** Let X be a Hilbertian random variable defined on $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}$ and let Y be another random variables on the same probability space with values in the measure space $(\mathcal{Y}, \mathbb{E})$. Assume that X has first moment. The conditional expectation $E(X \mid Y)$ then exists and is $\sigma(Y)$-measurable by construction, so by Theorem 4.2.7 there exists a measurable function $\phi : \mathcal{Y} \to \mathcal{H}$. We define $E(X \mid Y = y) := \phi(y)$ and call this *the conditional expectation of X given $Y = y$.*

As a natural generalization of the conditional covariance for real-valued random variables, we can define a conditional cross-covariance as below.

Definition 4.2.9 (Conditional cross-covariance)**.** Let X and Y be Hilbertian random variables defined on a common probability space $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively. Let $\mathbb{D} \subseteq \mathbb{F}$ be another σ-algebra. Assume that X and Y have finite second moment. We define the *conditional cross-covariance operator of X and Y given $\mathbb{D}$* by

$$\operatorname{Cov}(X, Y \mid \mathbb{D}) = E\left((Y - E(Y \mid \mathbb{D})) \odot_Y (X - E(X \mid \mathbb{D})) \mid \mathbb{D}\right).$$

We can rewrite this in a similar way as done for the cross-covariance.

Theorem 4.2.10 (Alternative expression for the conditional cross-covariance)**.** Let X and Y be Hilbertian random variables defined on a common probability space $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively. Let $\mathbb{D} \subseteq \mathbb{F}$ be another σ-algebra. Assume that X and Y have second moment. We can write the conditional cross-covariance as

$$\operatorname{Cov}(X, Y \mid \mathbb{D}) = E(Y \odot_Y X \mid \mathbb{D}) - E(Y \mid \mathbb{D}) \odot_Y E(X \mid \mathbb{D})$$

Proof.
Using linearity of the outer product in the definition of the conditional cross-covariance yields

$$\begin{aligned}\operatorname{Cov}(X, Y \mid \mathbb{D}) = {} & E(Y \odot_Y X \mid \mathbb{D}) - E(Y \odot_Y E(X \mid \mathbb{D}) \mid \mathbb{D}) \\ & - E(E(Y \mid \mathbb{D}) \odot_Y X \mid \mathbb{D}) + E(E(Y \mid \mathbb{D}) \odot_Y E(X \mid \mathbb{D}) \mid \mathbb{D}).\end{aligned}$$

The final term is equal to $E(Y \mid \mathbb{D}) \odot_Y E(X \mid \mathbb{D})$ since the outer product is continuous and thus measurable and therefore preserves the $\mathbb{D}$-measurability of the two conditional expectations. If we can show that both of the middle terms equal $-E(Y \mid \mathbb{D}) \odot_Y E(X \mid \mathbb{D})$, we will be done. This follows immediately from Theorem 4.2.5. □

We will later employ this conditional cross-covariance as the basis of the Hilbertian GCM, since it shares the following crucial property with the real-valued version.

Theorem 4.2.11 (Conditional cross-covariance of conditionally independent variables)**.** Let X and Y be Hilbertian random variables defined on a common probability space $(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively. Let $\mathbb{D} \subseteq \mathbb{F}$ be another σ-algebra. Assume that X and Y have second moment. Then if $X \perp\!\!\!\perp Y \mid \mathbb{D}$, we have $\mathrm{Cov}(X, Y \mid \mathbb{D}) = 0$.

Proof.
We show that $E(Y \odot_Y X \mid \mathbb{D}) = E(Y \mid \mathbb{D}) \odot_Y E(X \mid \mathbb{D})$ by showing that they perform the same operation on all $h \in \mathcal{H}_Y$. Taking $h \in \mathcal{H}_Y$ and an orthonormal basis $(e_n)_{n\in\mathbb{N}}$, we note that

$$\begin{aligned} E(Y \odot_Y X \mid \mathbb{D})h = E((Y \odot_Y X)h \mid \mathbb{D}) &= E(\langle h, Y\rangle X \mid \mathbb{D}) \\ &= E\left(\sum_{i=1}^{\infty} \langle h, Y\rangle\langle X, e_i\rangle e_i \mid \mathbb{D}\right) = \sum_{i=1}^{\infty} E(\langle h, Y\rangle\langle X, e_i\rangle \mid \mathbb{D}) e_i, \end{aligned}$$

by Theorem 4.2.4. We know that functions of random variables inherit conditional independence by Theorem 2.1.10, so $\langle X, e_i\rangle$ and $\langle h, Y\rangle$ are conditionally independent for all $i \in \mathbb{N}$. They are also integrable by assumption so their conditional expectation factorizes by Theorem 2.1.12. Therefore we get

$$\begin{aligned} &\sum_{i=1}^{\infty} E(\langle h, Y\rangle\langle X, e_i\rangle \mid \mathbb{D}) e_i = E(\langle h, Y\rangle \mid \mathbb{D}) \sum_{i=1}^{\infty} E(\langle X, e_i\rangle \mid \mathbb{D}) e_i \\ &= \langle h, E(Y \mid \mathbb{D})\rangle E\left(\sum_{i=1}^{\infty} \langle X, e_i\rangle e_i \mid \mathbb{D}\right) = (E(Y \mid \mathbb{D}) \odot_Y E(X \mid \mathbb{D}))h \end{aligned}$$

by various properties of the conditional expectation, thus proving that the conditional cross-covariance is zero as desired. □

Recall that to construct the univariate GCM, we saw that the product of the residuals of conditionally independent random variables had mean zero. We will now show that the same is true for the residuals of Hilbertian random variables under the outer product.

Theorem 4.2.12 (Product of residuals of conditionally independent Hilbertian variables is zero)**.** Let X and Y be Hilbertian random variables defined on a common probability space

$(\Omega, \mathbb{F}, P)$ with values in $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively. Let $\mathbb{D} \subseteq \mathbb{F}$ be another σ-algebra. Assume that X and Y have second moment.

Define the residuals $\varepsilon = X - E(X \mid \mathbb{D})$ and $\xi = Y - E(Y \mid \mathbb{D})$.

Then if $X \perp\!\!\!\perp Y \mid \mathbb{D}$, we have $E(\xi \odot_Y \varepsilon) = 0$.

Proof.
Note that by the tower property, it is sufficient to show that $E(\xi \odot_Y \varepsilon \mid \mathbb{D}) = 0$. Now by definition of the conditional cross-covariance and Theorem 4.2.11 we are done, since

$$E(\xi \odot_Y \varepsilon \mid \mathbb{D}) = E\left([Y - E(Y \mid \mathbb{D})] \odot_Y [X - E(X \mid \mathbb{D})] \mid \mathbb{D}\right) = \operatorname{Cov}(X, Y \mid \mathbb{D}).$$

□

4.3 Hilbertian estimation of moments and linear models

In this section we will discuss estimation of means and covariance for Hilbertian random variables. Then we will generalize the canonical linear model on Euclidean spaces to linear models for Hilbertian random variables. This will include the usual linear models as a special case. We will describe the necessary theoretical assumptions to ensure that both the model and estimation is well-defined, how to estimate in this framework and we will give a bound on the mean-squared prediction error using this estimator.

In the context of statistics it is of great importance whether we can estimate various properties of a distribution consistently. We will throughout assume that we are given n i.i.d observations of some Hilbertian random variable X. It is seen immediately from the law of large numbers that we can estimate the mean of X consistently, assuming it exists. For covariance operators the question is more subtle. Recall that the covariance was defined as an integral over the space of Hilbert-Schmidt operators, which is a Hilbert space, so we could once again apply the law of large numbers to note that covariances are estimated consistently. By this we mean that the estimates converge to the true covariance operator in Hilbert-Schmidt norm.

However we also noted that covariance operators are trace-class operators and a natural question becomes whether the estimates converge in trace norm to the true covariance operator. This is not at all obvious and follows from another version of the law of large numbers in Banach spaces. We will not go into the technical details of how to work with random variables on Banach spaces but much of the theory developed in the previous sections still holds, in particular the modes of convergence given in Definition 4.1.18 also hold in Banach spaces. For more on random variables on Banach Spaces see [16] or [28].

Theorem 4.3.1 (Law of Large Numbers in Banach spaces)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables with values in a Banach space $\mathcal{B}$. Assume that they are independent and identically distributed and that they have first moment. Denote their common mean by μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu.$$

For a proof of 4.3.1, see [3] Theorem 2.4. Using this theorem we can get that covariance estimation is also consistent in trace norm.

Corollary 4.3.2 (Consistency of covariance estimation)**.** Let $(X_n)_{n\in\mathbb{N}}$ be an i.i.d. sequence of Hilbertian random variables with second moment. Let $\mathscr{C}$ denote the covariance operator of the common distribution and define the empirical estimate of the covariance operator as

$$\widehat{\mathscr{C}_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \odot X_i.$$

Then

$$\|\widehat{\mathscr{C}_n} - \mathscr{C}\|_{TR} \overset{a.s.}{\to} 0.$$

We will now proceed to define a regression method for general Hilbert spaces. We follow the exposition given in [6], that is a generalization of the method given in [29]. This is a generalization of the finite-dimensional linear models and also of the functional linear model with scalar response given in Theorem 2.3.4. The possible infinite-dimensional nature of these spaces makes estimation a tricky problem and we will not go into the full details here. We give an account of this method since we will apply it in the empirical investigation of the upcoming conditional independence test on Hilbert spaces.

Definition 4.3.3 (Hilbertian linear model)**.** Let X and Y be Hilbertian random variables with values in $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively and let $\mathscr{S} \in \mathfrak{B}(\mathcal{H}_X, \mathcal{H}_Y)$. Let furthermore ε be a Hilbertian random variable with values in $\mathcal{H}_Y$ and $\varepsilon \perp\!\!\!\perp X$. Assume that

$$Y = \mathscr{S}X + \varepsilon.$$

The statistical model with sample space $\mathcal{H}_X \times \mathcal{H}_Y$, σ-algebra $\mathbb{B}(\mathcal{H}_X \times \mathcal{H}_Y)$ and set of distributions satisfying the relation above is the *Hilbertian linear model*.

The above assumptions are the minimal assumptions required for defining the model but we will need to impose further assumptions to construct an estimate of $\mathscr{S}$ and further still when considering the mean square prediction error of a new observation.

From now on we assume that

1. X and Y are centered, i.e. they have mean zero.
2. ε has finite second moment and covariance operator $\Gamma_\varepsilon = \mathrm{Cov}(\varepsilon)$.
3. $\mathscr{S}$ is a Hilbert-Schmidt operator.
4. X has moment of all orders and has injective covariance operator $\Gamma = \mathrm{Cov}(X)$ with convex eigenvalues from some point on, i.e. if $(\lambda_n)_{n\in\mathbb{N}}$ is the sequence of eigenvalues of Γ, the function that for $j \in \mathbb{N}$ maps j to λ_j, continuously interpolated on $\mathbb{R}^+$, is convex from some point on, i.e. there exists some $N \in \mathbb{N}$ so that the function is continuous for $x > N$.
5. If we let $(e_j)_{j\in\mathbb{N}}$ denote a basis of eigenvectors of Γ, we assume that there exists a constant b so that for all $k, j \in \mathbb{N}$ we have

$$E(|\langle X, e_j\rangle|^k) \leqslant \frac{k!}{2} b^{k-2} E(\langle X, e_j\rangle^2).$$

These assumptions are required to ensure that the model is identifiable (injectivity of Γ) and that estimation is well-behaved. For the full proofs and details regarding these assumptions, see [6]. If X is a Gaussian on $\mathcal{H}_X$ with injective covariance, the assumptions on Γ and X will be fulfilled.

We assume that we are given n i.i.d. samples from the model, $(X_i, Y_i)_{1\leqslant i\leqslant n}$ and attempt to estimate $\mathscr{S}(X^*)$ where X^* is a new independent observation from the model. Note that in practice it matters whether we're interested in $\mathscr{S}$ or $\mathscr{S}(X^*)$ when tuning the estimation process. More regularization is needed when estimating $\mathscr{S}$ than when estimating $\mathscr{S}(X^*)$ for reasons discussed in [4].

Definition 4.3.4 (Estimation in the Hilbertian linear model)**.** Continuing from Definition 4.3.3, we define

$$\Delta = \mathrm{Cov}(Y, X)$$

and empirical counterparts

$$\Gamma_n = \frac{1}{n}\sum_{i=1}^n X_i \odot_X X_i, \quad \text{and} \quad \Delta_n = \frac{1}{n}\sum_{i=1}^n X_i \odot_X Y_i.$$

Then letting $(k_n)_{n\in\mathbb{N}}$ be a sequence of natural numbers diverging and $(\widehat{\lambda}_j)_{j\in\mathbb{N}}$ and $(\widehat{e}_j)_{j\in\mathbb{N}}$ be the estimated eigenvalues and -vectors of Γ_n, we define

$$\Gamma_n^\dagger = \sum_{j=1}^{k_n} \widehat{\lambda}_j^{-1}(e_j \odot_X e_j)$$

and the estimator of $\mathscr{S}$

$$\widehat{\mathscr{S}_n} = \Delta_n \Gamma_n^\dagger.$$

To motivate the estimator given above, consider that by the method of moments, we have the equation

$$E(X \odot_X Y) = E(X \odot_X \mathscr{S}(X)) + E(X \odot_X \varepsilon).$$

The second term on the right hand side is zero, since we have assumed that $X \perp\!\!\!\perp \varepsilon$. The first term becomes $E(\mathscr{S}\Gamma)$ by linearity of $\mathscr{S}$, so in total we get

$$\Delta = \mathscr{S}\Gamma.$$

If Γ was invertible, we would simply multiply either side by the inverse and have an estimator but Γ is a covariance operator and thus compact and therefore has no inverse. The usage of $\Gamma^\dagger$ is a way around this problem, that is common in inverse problems.

The following result is a combination of Theorem 2 and parts of the proof of Theorem 9 in [6].

Theorem 4.3.5 (Prediction error in Hilbertian linear model)**.** Consider the setup given in Definition 4.3.3 and Definition 4.3.4. Letting $(\lambda_j, e_j)_{j\in\mathbb{N}}$ denote the eigenvalues and -vectors of Γ, we define

$$\gamma_k = \sup_{j \geqslant k}\{j \log j \|\mathscr{S}(e_j)\|\sqrt{\lambda_j}\}.$$

Assume that $(k_n \log k_n)^2/n \to 0$ and assume further that there exists some $N \in \mathbb{N}$ so that for all $n \geqslant N$, we have

$$\gamma_{k_n} \leqslant \frac{(k_n \log k_n)^2}{n},$$

then

$$\sqrt{n} E\|\widehat{\mathscr{S}_n}(X^*) - \mathscr{S}(X^*)\|^2 \to 0.$$

Proof.
Theorem 2 in [6] states that for any k, we have

$$\sqrt{n} E\|\widehat{\mathscr{S}_n}(X^*) - \mathscr{S}(X^*)\|^2 \leqslant \sigma_\varepsilon^2 \frac{k}{\sqrt{n}} + \sqrt{n} \sum_{j=k+1}^{\infty} \lambda_j \|\mathscr{S}(e_j)\|^2 + C_1 \|\mathscr{S}\|_{HS} \lambda_k \frac{k^2}{\sqrt{n}} + \frac{C_2}{\sqrt{n}} \frac{(k \log k)^2}{n},$$

where $\sigma_\varepsilon^2 = \|\Gamma_\varepsilon\|_{TR}$ and C_1 and C_2 are constants that do not depend on $\mathscr{S}$, k or n. It is obvious from the assumption that $(k_n \log k_n)^2/n \to 0$ that the fourth term goes to zero. Noting that this also implies $(k_n \log k_n)/\sqrt{n} \to 0$, which in turn implies $k_n/\sqrt{n} \to 0$, the first term also goes to zero. The third term goes to zero by noting that $\lambda_{k_n} k_n \to 0$, since $(\lambda_j)_{j\in\mathbb{N}}$ are summable and again by the previous argument.

The second term is the tricky one to deal with and arguments are given in the proof of Theorem 9 in [6], that show that using the assumption on γ_{k_n}, we have that

$$\frac{n}{k} \sum_{j=k+1}^{\infty} \lambda_j \|\mathscr{S}(e_j)\|^2 \to 0,$$

which implies that the second term also goes to zero, proving the result. □

The assumptions in Theorem 4.3.5 are mainly there to ensure that $(k_n \log k_n)^2/n$ goes to zero at the correct rate. We will apply this theorem in the next chapter where we extend the GCM to Hilbert spaces.

GCM in Hilbert spaces

In this chapter we extend the GCM to data with values in Hilbert spaces.

5.1 Definition and properties of the GHSCM

We have now developed the tools required to extend the GCM to Hilbertian random variables. We will consider X and Y Hilbertian random variables defined on $(\Omega, \mathbb{F}, P)$ with values in two Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ of possibly infinite dimension. We further have a random variable Z defined on the same probability space with values in a space $\mathcal{Z}$ that is only used for prediction i.e. conditioning. The requirements on Z are identical to the univariate GCM; that we can construct a sub-σ-algebra of $\mathbb{F}$ that we can use for conditioning and that we have a regression method with a suitable rate of convergence when regressing X on Z and Y on Z. The regression requirement will typically be the limiting one but if we assume that $\mathcal{Z}$ is a third Hilbert space, we have the regression methods explained in Section 4.3. We will thus simply assume that $\mathcal{Z}$ is in a measurable space $(\mathcal{Z}, \mathbb{G})$.

The Generalised Hilbert Space Covariance Measure (GHSCM) retains the gist of the GCM for univariate random variables but is considerably more complicated due to the fact that we cannot normalize the asymptotic limit distribution like in the univariate case. In the univariate case we considered a covariance estimator of residuals (which is an operator when $\mathbb{R}$ is viewed as a Hilbert space) that we argued had a limiting normal distribution with some variance in Theorem 2.3.2. The test-statistic was constructed to whiten the asymptotic distribution so that the Gaussian limit was always standard. On an infinite dimensional Hilbert space this is impossible, since there is no standard Gaussian in infinite dimensions.

We will proceed in a manner similar to the univariate case but without normalization the asymptotic distribution of our test statistic becomes quite different.

Definition 5.1.1 (GHSCM test statistic)**.** Let X and Y be random variables with values in two Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ with inner products $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_Y$ and norms $\|\cdot\|_X$

and $\|\cdot\|_Y$ respectively. Let Z be a random variable with values in measurable space $(\mathcal{Z}, \mathbb{G})$. Consider the statistical model of all joint distributions $\mathcal{H}_X \times \mathcal{H}_Y \times \mathcal{Z}$, i.e.

$$\mathcal{P} = \{\nu \text{ probability measure on } (\mathcal{H}_X \times \mathcal{H}_Y \times \mathcal{Z}, \mathbb{B}(\mathcal{H}_X) \otimes \mathbb{B}(\mathcal{H}_Y) \otimes \mathbb{G}\}.$$

Consider the hypothesis $X \perp\!\!\!\perp Y \mid Z$ with corresponding subset of probability measures $\mathcal{P}_0$. For every $\nu \in \mathcal{P}$, we can write

$$X = \underbrace{E_\nu(X \mid Z)}_{f_\nu(Z)} + \underbrace{X - E_\nu(X \mid Z)}_{\varepsilon_\nu},$$

i.e. $f_\nu(z) = E_\nu(X \mid Z = z)$ and similarly

$$Y = \underbrace{E_\nu(Y \mid Z)}_{g_\nu(Z)} + \underbrace{Y - E_\nu(Y \mid Z)}_{\xi_\nu}.$$

Let $(x, y, z)^{(n)} \in (\mathcal{H}^2 \times \mathcal{Z})^n$ be a sample of size n from the model and let $\widehat{f}^{(n)}$ and $\widehat{g}^{(n)}$ denote estimates of f and g based on the sample. For $i = 1, \ldots, n$ define

$$\mathscr{R}_i^{(n)} = (y_i - \widehat{g}^{(n)}(z_i)) \odot_Y (x_i - \widehat{f}^{(n)}(z_i))$$

and define

$$T_n = \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathscr{R}_i^{(n)} \right\|_{HS}^2,$$

where $\|\cdot\|_{HS}$ is the Hilbert-Schmidt norm on $\mathfrak{B}_{HS}(\mathcal{H}_Y, \mathcal{H}_X)$. This is the *Generalised Hilbert Space Covariance Measure* (GHSCM) test statistic.

Theorem 5.1.2 (Asymptotic distribution of GHSCM test statistic)**.** Continuing from Definition 5.1.1, we define for each $\nu \in \mathcal{P}$

$$u_\nu(z) = E_\nu\left(\|\varepsilon_\nu\|^2 \mid Z = z\right), \quad v_\nu(z) = E_\nu\left(\|\xi_\nu\|^2 \mid Z = z\right).$$

We further define the mean-squared prediction error and weighted mean-squared prediction error for f

$$M_{\nu,n}^f = \frac{1}{n} \sum_{i=1}^n \left\| f_\nu(z_i) - \widehat{f}^{(n)}(z_i) \right\|^2 \quad \text{and} \quad \tilde{M}_{\nu,n}^f = \frac{1}{n} \sum_{i=1}^n \left\| f_\nu(z_i) - \widehat{f}^{(n)}(z_i) \right\|^2 v_\nu(z_i),$$

and

$$M_{\nu,n}^g = \frac{1}{n} \sum_{i=1}^n \left\| g_\nu(z_i) - \widehat{g}^{(n)}(z_i) \right\|^2 \quad \text{and} \quad \tilde{M}_{\nu,n}^g = \frac{1}{n} \sum_{i=1}^n \left\| g_\nu(z_i) - \widehat{g}^{(n)}(z_i) \right\|^2 u_\nu(z_i),$$

for g.

Assume that for each $\nu \in \mathcal{P}_0$, $nM^f_{\nu,n}M^g_{\nu,n} \xrightarrow{P} 0$, $\tilde{M}^f_{\nu,n} \xrightarrow{P} 0$, $\tilde{M}^g_{\nu,n} \xrightarrow{P} 0$ and $0 < E_\nu\left(\|\varepsilon_\nu\|^2\|\xi_\nu\|^2\right) < \infty$ then for every $\nu \in \mathcal{P}_0$, we have

$$T_n \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty} \lambda_i W_i^2,$$

where $(W_n)_{n\in\mathbb{N}}$ is an i.i.d sequence of standard normal variables and $(\lambda_i)_{i=1}^{\infty}$ is the non-increasing sequence of eigenvalues for $\mathrm{Cov}(\xi \odot_Y \varepsilon)$.

Proof.
In the following we suppress dependence on $\nu \in \mathcal{P}_0$, since all the calculations are identical for all measures.

Note that if we can show that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathscr{R}_i^{(n)}$ converges to a Gaussian with mean zero and covariance operator $\mathrm{Cov}(\xi \odot_Y \varepsilon)$, we will be done according to Theorem 4.1.15.

By arguments similar to the univariate GCM proof and linearity of the outer product, we get

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathscr{R}_i^{(n)} = &\underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \odot_Y \varepsilon_i}_{U_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n (g(z_i) - \widehat{g}^{(n)}(z_i)) \odot_Y (f(z_i) - \widehat{f}^{(n)}(z_i))}_{a_n} \\ &+ \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \odot_Y (f(z_i) - \widehat{f}^{(n)}(z_i))}_{b_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n (g(z_i) - \widehat{g}^{(n)}(z_i)) \odot_Y \varepsilon_i}_{c_n}.\end{aligned}$$

Since $\nu \in \mathcal{P}_0$, Theorem 4.2.12 yields that the sequence of Hilbert-Schmidt operators $(\xi_i \odot_Y \varepsilon_i)_{i\in\mathbb{N}}$ has mean zero and by assumption they are i.i.d with finite variance, thus the Hilbertian CLT gives that $U_n \xrightarrow{\mathcal{D}} G$, where G is a Gaussian with mean zero and covariance operator equal to the covariance of $\xi \odot_Y \varepsilon$. By Slutsky's theorem if a_n, b_n and c_n all converge to 0 in probability, we will be done. We will establish this by looking at the square of the Hilbert-Schmidt norm of the sequences, since convergence of the squared norms to 0 implies convergence of the sequences to 0.

Note that multiplicativity and sub-additivity of norms, Theorem 3.2.27 and Cauchy-Schwarz,

we get

$$\begin{aligned}
\|a_n\|_{HS}^2 &= \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (g(z_i) - \widehat{g}^{(n)}(z_i)) \odot_Y (f(z_i) - \widehat{f}^{(n)}(z_i)) \right\|_{HS}^2 \\
&\leqslant \frac{1}{n} \sum_{i=1}^{\infty} \|(g(z_i) - \widehat{g}^{(n)}(z_i)) \odot_Y (f(z_i) - \widehat{f}^{(n)}(z_i))\|_{HS}^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \|g(z_i) - \widehat{g}^{(n)}(z_i)\|_Y^2 \|(f(z_i) - \widehat{f}^{(n)}(z_i))\|_X^2 \\
&\leqslant \frac{1}{n} \sum_{i=1}^{n} \|g(z_i) - \widehat{g}^{(n)}(z_i)\|_Y^2 \sum_{i=1}^{n} \|(f(z_i) - \widehat{f}^{(n)}(z_i))\|_X^2 = n M_n^f M_n^g \xrightarrow{P} 0,
\end{aligned}$$

by assumption. To establish that $\|b_n\|_{HS}^2$ goes to 0, we will apply Lemma 2.1.28 and show that the conditional expectation given $X^{(n)}$ and $Z^{(n)}$ goes to zero, which will imply the desired result. We get by using the relationship between the Hilbert-Schmidt norm and inner product and by linearity of both that

$$\begin{aligned}
E(\|b_n\|_{HS}^2 \mid X^{(n)}, Z^{(n)}) &= \frac{1}{n} E\left(\left\| \sum_{i=1}^{n} \xi_i \odot_Y (f(z_i) - \widehat{f}^{(n)}(z_i)) \right\|_{HS}^2 \mid X^{(n)}, Z^{(n)} \right) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} E\left(\langle \xi_i \odot_Y (f(z_i) - \widehat{f}^{(n)}(z_i)), \xi_j \odot_Y (f(z_j) - \widehat{f}^{(n)}(z_j)) \rangle_{HS} \mid X^{(n)}, Z^{(n)} \right) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} E\left(\langle f(z_i) - \widehat{f}^{(n)}(z_i), f(z_j) - \widehat{f}^{(n)}(z_j) \rangle_X \langle \xi_i, \xi_j \rangle_Y \mid X^{(n)}, Z^{(n)} \right) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \langle f(z_i) - \widehat{f}^{(n)}(z_i), f(z_j) - \widehat{f}^{(n)}(z_j) \rangle_X E\left(\langle \xi_i, \xi_j \rangle_Y \mid X^{(n)}, Z^{(n)} \right),
\end{aligned}$$

where the second to last equality to due to Theorem 3.2.27 and the last equality holds since the terms involving $f(z_i) - \widehat{f}^{(n)}(z_i)$ are measurable wrt. the σ-algebra generated by $X^{(n)}$ and $Z^{(n)}$. $\langle \xi_i, \xi_j \rangle$ only depends on Z_i and Z_j of the conditioning variables, so we can omit the remaining variables form the conditioning expression.

For $i \neq j$, by using that $E(Y_i \mid Z_i) = E(Y_i \mid Z_i, Z_j)$ since Z_j is independent of (Y_i, Z_i) and Theorem 4.2.5, we get

$$\begin{aligned}
E\left(\langle \xi_i, \xi_j \rangle_Y \mid X^{(n)}, Z^{(n)} \right) &= E[\langle Y_i, Y_j \rangle_Y - \langle Y_i, E(Y_j \mid Z_j) \rangle_Y \\
&\quad - \langle E(Y_i \mid Z_i), Y_j \rangle_Y + \langle E(Y_i \mid Z_i), E(Y_j \mid Z_j) \rangle_Y \mid Z_i, Z_j] \\
&= E(\langle Y_i, Y_j \rangle_Y \mid Z_i, Z_j) - \langle E(Y_i \mid Z_i, Z_j), E(Y_j \mid Z_i, Z_j) \rangle_Y.
\end{aligned}$$

We will show that this is zero, by first recalling that by assumption $(Y_i, Z_i) \perp\!\!\!\perp (Y_j, Z_j)$, so applying weak union and symmetry from Theorem 2.1.9, we get $Y_i \perp\!\!\!\perp Y_j \mid (Z_i, Z_j)$. Take

now some orthonormal basis for $\mathcal{H}_Y$, $(e_k)_{k\in\mathbb{N}}$ and note that

$$E(\langle Y_i, Y_j\rangle_Y \mid Z_i, Z_j) = E\left(\sum_{k=1}^{\infty}\langle Y_i, e_k\rangle_Y\langle Y_j, e_k\rangle_Y \mid Z_i, Z_j\right) = \sum_{k=1}^{\infty} E\left(\langle Y_i, e_k\rangle_Y\langle Y_j, e_k\rangle_Y \mid Z_i, Z_j\right).$$

Note that for all k, $\langle Y_i, e_k\rangle_Y$ and $\langle Y_j, e_k\rangle_Y$ are conditionally independent given (Z_i, Z_j), so by Theorem 2.1.12, the conditional expectation factorizes and we get

$$\begin{aligned}&\sum_{k=1}^{\infty} E\left(\langle Y_i, e_k\rangle_Y\langle Y_j, e_k\rangle_Y \mid Z_i, Z_j\right) = \sum_{k=1}^{\infty} E(\langle Y_i, e_k\rangle_Y \mid Z_i, Z_j)E(\langle Y_j, e_k\rangle_Y \mid Z_i, Z_j)\\ &= \sum_{k=1}^{\infty}\langle E(Y_i \mid Z_i, Z_j), e_k\rangle_Y\langle E(Y_j \mid Z_i, Z_j), e_k\rangle_Y = \langle E(Y_i \mid Z_i, Z_j), E(Y_j \mid Z_i, Z_j)\rangle_Y,\end{aligned}$$

by various rules for manipulating the conditional expectation.

We can thus omit all terms from the sum where $i \neq j$ and therefore

$$E(\|b_n\|_{HS}^2 \mid X^{(n)}, Z^{(n)}) = \frac{1}{n}\sum_{i=1}^{n}\|f(z_i) - \widehat{f}^{(n)}(z_i)\|_X^2 E\left(\|\xi_i\|_Y^2 \mid Z_i\right) = \tilde{M}_n^f \xrightarrow{P} 0,$$

by assumption. An analogous argument can be repeated for c_n, thus proving the desired result. □

The test statistic that we have constructed above differs from the one employed by the regular GCM. First of all the limiting distribution depends on the underlying distribution through the eigenvalues of the covariance of $\xi \odot_Y \varepsilon$, which could vary greatly depending on the underlying distributions. Further, the limiting distribution has no known closed form expressions associated with it – we have no density, distribution or quantile functions to employ.

We can estimate the sequence of eigenvalues consistently by Theorem 4.3.2 and since the sequence of eigenvalues tends to 0 sufficiently fast to ensure that the sequence is in ℓ^1, we could approximate the infinite sum of weighted chi squares with a truncated sum and calculate quantiles through bootstrapping. Let us formalize the resulting test:

Definition 5.1.3 (GHSCM)**.** Continuing from Theorem 5.1.2, we denote by $(\widehat{\lambda}_n)_{n\in\mathbb{N}}$ the sequence of estimates of λ, the eigenvalues of $\mathrm{Cov}(\xi \odot_Y \varepsilon)$. The sequence $\widehat{\lambda}_n$ is finite for every n, so let k_n denote the number of non-zero estimates and let $\widehat{\lambda}_n(i)$ denote the i'th element of the sequence n'th sequence.

Let now $\alpha \in (0,1)$ and $(b_n)_{n\in\mathbb{N}}$ a sequence of natural numbers diverging and denote by $\widehat{q}^{(l)}_{\widehat{\lambda}_n}$ and $\widehat{q}^{(u)}_{\widehat{\lambda}_n}$ the empirical $\alpha/2$ and $1-\alpha/2$ quantiles respectively, obtained through bootstrapping the distribution of $V_n = \sum_{i=1}^{k_n}\widehat{\lambda}_n(i)W_i^2$ using b_n samples.

Then the *Generalised Hilbert Space Covariance Measure* is the test $(\psi_n)_{n\in\mathbb{N}}$ given by

$$\psi_n((x,y,z)^n) = \begin{cases} 0 & \text{if } T_n \in [\widehat{q}^{(l)}_{\hat{\lambda}_n}, \widehat{q}^{(u)}_{\hat{\lambda}_n}] \\ 1 & \text{otherwise.} \end{cases}$$

The test above is constructed in a manner similar to the univariate GCM, that was constructed based on an asymptotic test statistic. However T_n is strictly speaking not a test statistic, since its distribution depends on the underlying probability measure. While we did find an expression for the asymptotic distribution of T_n, this is not sufficient to prove asymptotic pointwise level. To do so, we will need a few preliminary observations and lemmas.

Lemma 5.1.4. Let $(W_n)_{n\in\mathbb{N}}$ be an i.i.d sequence of standard Gaussian random variables, let $(\lambda_n)_{n\in\mathbb{N}}$ be a random sequence of absolutely summable, positive sequences independent of $(W_n)_{n\in\mathbb{N}}$. Let further λ be a static absolutely summable, positive sequence. Then if

$$\sum_{i=1}^{\infty} |\lambda_n(i) - \lambda(i)| \xrightarrow{a.s.} 0,$$

we have

$$\sum_{i=1}^{\infty} \lambda_n(i) W_i^2 \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty} \lambda(i) W_i^2.$$

Proof.
Note that if we can show

$$\sum_{i=1}^{\infty} \lambda_n(i) W_i^2 - \sum_{i=1}^{\infty} \lambda(i) W_i^2 \xrightarrow{P} 0,$$

by Slutsky's theorem, we will be done. To that end let $\varepsilon > 0$ be given and note that by the triangle inequality and Markov's inequality, we have

$$P\left(\left|\sum_{i=1}^{\infty} \lambda_n(i) W_i^2 - \sum_{i=1}^{\infty} \lambda(i) W_i^2\right| \geqslant \varepsilon\right) \leqslant \frac{E\left(\sum_{i=1}^{\infty} |\lambda_n(i) - \lambda(i)|\right)}{\varepsilon},$$

since $E(W_i^2) = 1$ and $(W_n)_{n\in\mathbb{N}}$ is independent of $(\lambda_n)_{n\in\mathbb{N}}$. If we can show that the integral converges to 0, we will be done. Note that the Dominated Converge Theorem yields the desired result if the sequence is dominated. By the triangle and reverse triangle inequalities, note that

$$\left|\sum_{i=1}^{\infty} |\lambda_n(i)| - \sum_{i=1}^{\infty} |\lambda(i)|\right| \leqslant \sum_{i=1}^{\infty} ||\lambda_n(i)| - |\lambda(i)|| \leqslant \sum_{i=1}^{\infty} |\lambda_n(i) - \lambda(i)|,$$

which goes to zero almost surely by assumption, so $\sum_{i=1}^{\infty} |\lambda_n(i)| \to \sum_{i=1}^{\infty} |\lambda(i)|$ and thus we can choose N parrying $\varepsilon = 1$, to get

$$\sum_{i=1}^{\infty} |\lambda_n(i)| \leqslant \sum_{i=1}^{\infty} |\lambda(i)| + 1,$$

for all $n \geqslant N$. Set

$$C = \max_{n<N} \left\{ \sum_{i=1}^{\infty} |\lambda_n(i)| \right\},$$

and note that by the triangle inequality

$$\sum_{i=1}^{\infty} |\lambda_n(i) - \lambda(i)| \leqslant \sum_{i=1}^{\infty} |\lambda_n(i)| + \sum_{i=1}^{\infty} |\lambda(i)| \leqslant 2 \sum_{i=1}^{\infty} |\lambda(i)| + C + 1,$$

thus proving the statement. □

Lemma 5.1.5. Let $f_n : \mathbb{R} \to \mathbb{R}$ be a sequence of non-decreasing functions converging uniformly to a strictly increasing limit function $f : \mathbb{R} \to \mathbb{R}$. Then the sequence of generalised inverses f_n^- converges pointwise to f^-.

Proof.
Consider a fixed $y_0 \in \mathbb{R}$ and let $\varepsilon > 0$ be given. We need to show that there exists $N \in \mathbb{N}$ so that for all $n \geqslant N$, we have

$$|f_n^-(y_0) - f^-(y_0)| \leqslant \varepsilon.$$

Let $x_0 = f^-(y_0)$ and note that since f is strictly increasing, we can find $\delta > 0$, so that

$$f(x_0 - \varepsilon) + \delta < \underbrace{y_0}_{=f(x_0)} < f(x_0 + \varepsilon) - \delta.$$

Now choose $N \in \mathbb{N}$ from the uniform convergence of f_n to f, such that $\sup_{x \in \mathbb{R}} |f_n(x) - f(x)| < \delta$. For $n \geqslant N$ we now have $f_n(x_0 - \varepsilon) < f(x_0 - \varepsilon) + \delta < y_0$ and $f_n(x_0 + \varepsilon) > f(x_0 + \varepsilon) - \delta > y_0$. So by applying f_n^- to either side we get

$$x_0 - \varepsilon < f_n^-(y_0) < x_0 + \varepsilon,$$

and therefore

$$|f_n^-(y_0) - f^-(y_0)| = |f_n^-(y_0) - x_0| < \varepsilon,$$

as desired. □

The lemma above in particular applies to distribution functions in the sense that if a sequence of distribution functions converges uniformly to the distribution function of a continuous distribution (which is thus strictly increasing), we have that the quantile functions converge pointwise.

We are now ready to prove that the GHSCM has asymptotic pointwise level.

Theorem 5.1.6 (Pointwise asymptotic level of GHSCM)**.** Continuing from Definition 5.1.3, under the assumptions of Theorem 5.1.2 the GHSCM has pointwise asymptotic level.

Proof.
If we can show, that for any $\nu \in \mathcal{P}_0$, we have

$$\lim_{n\to\infty} P_\nu(\psi_n = 1) = \alpha$$

we will be done (we will omit the ν from here on).

Note that

$$P(\psi_n = 1) = P\left(T_n \notin \left[\widehat{q}^{(l)}_{\widehat{\lambda}_n}, \widehat{q}^{(u)}_{\widehat{\lambda}_n}\right]\right) = 1 - F_{T_n}\left(\widehat{q}^{(u)}_{\widehat{\lambda}_n}\right) + F_{T_n}\left(\widehat{q}^{(l)}_{\widehat{\lambda}_n}\right)$$

where F_{T_n} denotes the CDF of T_n and $\widehat{q}^{(l)}_{\widehat{\lambda}_n}$ and $\widehat{q}^{(u)}_{\widehat{\lambda}_n}$ denote the estimates of the $\alpha/2$ and $1-\alpha/2$ quantiles of $\sum_{i=1}^{\infty} \lambda_i W_i^2$ using the bootstrap and the estimate $\widehat{\lambda}$. Letting F_λ denote the CDF of $\sum_{i=1}^{\infty} \lambda_i W_i^2$ and q_λ denote a true quantile, we will be done, if we can show that $F_{T_n}(\widehat{q}_{\widehat{\lambda}}) \to F_\lambda(q_\lambda)$ for any quantile.

By the triangle inequality, we can write

$$|F_{T_n}(\widehat{q}_{\widehat{\lambda}}) - F_\lambda(q_\lambda)| \leqslant |F_{T_n}(\widehat{q}_{\widehat{\lambda}}) - F_\lambda(\widehat{q}_{\widehat{\lambda}})| + |F_\lambda(\widehat{q}_{\widehat{\lambda}}) - F_\lambda(q_{\widehat{\lambda}})| + |F_\lambda(q_{\widehat{\lambda}}) - F_\lambda(q_\lambda)|.$$

The first term goes to zero by Theorem 5.1.2. For the second term, note that Glivenko-Cantelli (see the appendix, Theorem A.1.21) yields uniform convergence of the empirical distribution functions to the true distribution functions and Lemma 5.1.5 yields that the corresponding quantile functions converge, since $F_{\widehat{\lambda}}$ is continuous for any λ that is not the zero sequence. The term then goes to zero by continuity of F_λ. For the third term, Lemma 5.1.4 yields convergence of quantiles, since we know that $\widehat{\lambda}$ converges to λ in ℓ^1, since this is equivalent to convergence of the estimated covariance operator in trace norm, which follows from Corollary 4.3.2. The term then goes to zero by continuity of F_λ. □

The above theorem proves that we have in fact constructed an asymptotically valid test.

Unlike the univariate GCM, we do not have the necessary tools to prove a uniform asymptotic version of the GHSCM. To the best of the author's knowledge, no uniform limit theorems exist on Hilbert spaces. While some results exist about the uniform validity of the bootstrap (see [21] for instance), it is also unclear whether the eigenvalues of $\mathrm{Cov}(\xi \odot_Y \varepsilon)$ can be estimated uniformly. Thus it is doubtful whether we can state satisfying conditions to ensure uniform asymptotic level of the GHSCM.

Using the Hilbertian linear model, we can construct a complete example of a conditional independence test.

Theorem 5.1.7 (GHSCM using the Hilbertian linear model)**.** Let X, Y and Z be Hilbertian random variables in three possibly distinct Hilbert spaces $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$ respectively.

Assume that regressing X on Z and Y on Z both satisfy the assumptions in Theorem 4.3.5 and assume further that $u_\nu(z)$ and $v_\nu(z)$ in Theorem 5.1.2 are bounded by some $\sigma^2 > 0$. Then the GHSCM testing $X \perp\!\!\!\perp Y \mid Z$ has pointwise asymptotic level.

Proof.
We will only need to show that $\sqrt{n}M^f_{\nu,n}$ and $\sqrt{n}M^g_{\nu,n}$ go to zero in probability, since the remaining conditions then follow by the assumptions. This holds by Markov's inequality and Theorem 4.3.5. □

The author believes, that this is the first example of a conditional independence test for Hilbertian data (and thus also for functional data, since these can be viewed as random elements in a Hilbert space).

5.2 Empirical investigation of the GHSCM

In this section we will compare the performance of the GCM with the GHSCM through a simulation study. We will consider Hilbertian random variables in ℓ^2, the prototypical example of an infinite-dimensional Hilbert space and throughout let $(e_n)_{n\in\mathbb{N}}$ denote the standard basis in ℓ^2 as described in Example 3.1.7. We will simulate X, Y and Z as Hilbertian random variables in various ways as described later. The simulations will not be truly infinite-dimensional but we will instead simulate truncated versions of the variables. Throughout the study we only simulate the first 50 components of the infinite-dimensional variables considered.

The main purpose of this simulation study is to investigate situations where the real-valued GCM does not apply and where the GHSCM is more appropriate to use. Given n observations of infinite-dimensional Hilbertian random variables X, Y and Z we can perform a principal components analysis and retain the k components with the most variation in each of the variables. Denote these k principal components of the variables X', Y' and Z' respectively. Supposing that we have a regression method that regresses X' on Z' and Y' on Z' that satisfies the requirements for the GCM, we have a test with asymptotic level. However it is not clear whether we are able to detect all cases where conditional independence is not present. For instance, we could imagine that the dependence happens in the omitted parts of the variables and we would never be able to detect it. We cannot increase the number of principal components as n increases in the GCM, since then we no longer have a level guarantee. The GHSCM combined with the Hilbertian linear model allows us to retain asymptotic level while hopefully being able to detect more forms of dependence than the regular GCM.

In the simulation study we will consider four models. We consider two models, where we can theoretically justify the use of the GHSCM by Theorem 5.1.7. For both of these models we let ε_X, ε_Y and ε_Z be independent Hilbertian random variables with mean zero and covariance operator

$$\Gamma_\varepsilon = \sum_{n=1}^{\infty} \frac{1}{2^n}(e_n \odot e_n).$$

Let furthermore $\mathscr{S}_k$ be the k-shift operator, i.e. $\mathscr{S}_k((a_1, a_2, a_3, \dots)) = (a_k, a_{k+1}, a_{k+2}, \dots)$ and

$$\mathscr{A}_{k,c} = \left(\sum_{n=1}^{\infty} \frac{c}{n}(e_n \odot e_n) \right) \mathscr{S}_k.$$

We consider the models:

1. $$\begin{aligned} Z &= \varepsilon_Z \\ X &= \mathscr{A}_{1,1} Z + \varepsilon_X \\ Y &= \mathscr{A}_{1,1} Z + \varepsilon_Y. \end{aligned}$$

   In this model H_0 is true, i.e. $X \perp\!\!\!\perp Y \mid Z$.

2. $$\begin{aligned} Z &= \varepsilon_Z \\ X &= \mathscr{A}_{1,1} Z + \varepsilon_X \\ Y &= \mathscr{A}_{1,1} Z + \mathscr{A}_{5,5} X + \varepsilon_Y. \end{aligned}$$

   In this model H_0 is false, i.e. $X \not\perp\!\!\!\perp Y \mid Z$.

For these models Theorem 5.1.7 yields that we have asymptotic level if we choose k_n sensibly.

Theorem 5.2.1 (GHSCM in model 1 and 2 has pointwise asymptotic level)**.** Using $k_n = \lceil n^{2/5} \rceil$ in the estimation procedure of the Hilbertian linear model, the GHSCM in model 1 and 2 satisfy the conditions of Theorem 5.1.7 and thus we can test conditional independence with pointwise level.

Proof.
It is sufficient to show that the conditions are satisfied in model 2, since the regression X on Z in model is identical to both regressions in model 1.

Consider first regressing X on Z. It is clear that $\mathscr{A}_{k,c}$ is Hilbert-Schmidt since it is the composition of a Hilbert-Schmidt operator and a bounded operator. We also know that all

of the technical assumptions for the regression are fulfilled since Z is Gaussian and the noise is mean zero Gaussian. We only need to ensure that we choose k_n in a way satisfying that both $(k_n \log k_n)^2/n \to 0$ and that there exists some N so that $\gamma_{k_n} \leqslant (k_n \log k_n)^2/n$ for all $n \geqslant N$ as in Theorem 4.3.5.

Recall that $\gamma_k = \sup_{j \geqslant k}\{j \log j \|\mathscr{A}_{1,1}(e_j)\|\sqrt{\lambda_j}\}$ where (λ_j, e_j) is the j'th eigenvalue and -vector pair of the covariance of Z. By construction we have, $\sqrt{\lambda_j} = \sqrt{2^{-j}}$ and that the eigenvectors are the basis vectors $(e_n)_{n\in\mathbb{N}}$. We also get $\|\mathscr{A}_{1,1}(e_j)\| = \frac{1}{j-1}$ for $j > 1$ and 0 if $j = 1$. Thus $\gamma_1 = \gamma_2$ and for all $k > 1$ and we have

$$\gamma_k = \frac{k \log k}{(k-1)\sqrt{2^k}}.$$

Since

$$\frac{4}{25}\frac{\log(n)^2}{n^{1/5}} \leqslant \frac{\lceil n^{2/5}\rceil^2 \log(\lceil n^{2/5}\rceil)^2}{n} = \frac{(k_n \log k_n)^2}{n}$$

and

$$\gamma_{k_n} = \frac{\lceil n^{2/5}\rceil \log(\lceil n^{2/5}\rceil)}{(\lceil n^{2/5}\rceil - 1)\sqrt{2^{\lceil n^{2/5}\rceil}}} \leqslant \frac{(n^{2/5}+1)\log(n^{2/5}+1)}{(n^{2/5}-1)\sqrt{2^{n^{2/5}}}},$$

we will be done, if we can find N such that for all $n \geqslant N$, we have

$$\frac{(n^{2/5}+1)\log(n^{2/5}+1)}{(n^{2/5}-1)\sqrt{2^{n^{2/5}}}} \leqslant \frac{4}{25}\frac{\log(n)^2}{n^{1/5}}.$$

Such an N exists if

$$\frac{(n^{2/5}+1)\log(n^{2/5}+1)}{(n^{2/5}-1)\sqrt{2^{n^{2/5}}}}\frac{25}{4}\frac{n^{1/5}}{\log(n)^2} \to 0$$

as $n \to \infty$. Equivalently we can show that

$$\frac{(m+1)\log(m+1)}{(m-1)\sqrt{2^m}}\frac{25}{4}\frac{\sqrt{m}}{\log\left(m^{5/2}\right)^2} \to 0$$

as $m \to \infty$.

This is true since we can write

$$\frac{(m+1)\log(m+1)}{(m-1)\sqrt{2^m}}\frac{25}{4}\frac{\sqrt{m}}{\log\left(m^{5/2}\right)^2} = \frac{m+1}{m-1}\sqrt{\frac{m}{2^m}}\frac{\log(m+1)}{\log(m)^2}$$

where the first factor goes to 1, the second to 0 and the third to 0 by an application of L'Hôpital's rule.

When regressing Y on Z, note that we can write

$$Y = (\mathscr{A}_{1,1} + \mathscr{A}_{5,5}\mathscr{A}_{1,1})Z + \mathscr{A}_{5,5}\varepsilon_X + \varepsilon_Y.$$

The noise is again Gaussian by Theorem 4.1.12 and therefore satisfies the conditions of the regression. Z is still Gaussian and thus also satisfies the conditions of the regression. It remains to show the condition on γ_{k_n}. We can calculate γ_k as before except now defining $\mathscr{S} = \mathscr{A}_{1,1} + \mathscr{A}_{5,5}\mathscr{A}_{1,1}$ we have

$$\|\mathscr{S}(e_j)\| = \begin{cases} 0 & \text{if } j = 1 \\ \frac{1}{j-1} & \text{if } j \in \{2,3,4,5,6\} \\ \sqrt{\frac{1}{(j-1)^2} + \frac{25}{(j-1)^2(j-6)^2}} & \text{otherwise} \end{cases}$$

and thus $\gamma_1 = \gamma_2$ and for $k \in \{2,3,4,5,6\}$ γ_k is identical to when regressing X on Z and for $k > 6$

$$\gamma_k = \frac{k \log k}{\sqrt{2^k}} \sqrt{\frac{1}{(k-1)^2} + \frac{25}{(k-1)^2(k-6)^2}}.$$

We can now proceed with similar arguments as before. □

We also consider two further models where we do not have results guaranteeing the validity of the GHSCM. Let $\tilde{\varepsilon}_X$ and $\tilde{\varepsilon}_Y$ be independent mean zero Hilbertian random variables with covariance operator

$$\Gamma_{\tilde{\varepsilon}} = \sum_{n=1}^{5} (e_n \odot e_n) + \sum_{n=6}^{\infty} \frac{1}{2^{n-5}} (e_n \odot e_n).$$

We consider the models:

3.

$$\begin{aligned} X &= \tilde{\varepsilon}_X \\ Z &= \mathscr{A}_{1,1} X + \varepsilon_Z \\ Y &= \mathscr{A}_{1,1} Z + \tilde{\varepsilon}_Y. \end{aligned}$$

In this model H_0 is true, i.e. $X \perp\!\!\!\perp Y \mid Z$.

4.

$$\begin{aligned} X &= \tilde{\varepsilon}_X \\ Y &= \tilde{\varepsilon}_Y \\ Z &= \mathscr{A}_{5,5} X + \mathscr{A}_{5,5} Y + \varepsilon_Z \end{aligned}$$

In this model H_0 is false, i.e. $X \not\perp\!\!\!\perp Y \mid Z$.

It is unclear whether the mean square error of the Hilbertian linear model converges for model 3 and 4 as it does for model 1 and 2.

While this simulation study is done in ℓ^2, we could just as well visualise it in $L^2[0,1]$ since these spaces are congruent. In Figure 5.1 we see 3 realizations from each of the four models plotted in the Fourier basis of L^2. In Figure 5.2 we write each of the models as a graphical model to express the causal relations between the variables (for more on the use of graphical models to express conditional independence, see [19].)

In the simulation study we compare the GCM and the GHSCM. For the GCM we perform principal components analysis and retain the 5 components with most variation and employ linear least-squares regression as the regressor. To compute the GCM for multivariate data we construct the univariate GCM for each combination of components of the residual (yielding 25 components) and consider the sum of squares. The sum of squares is then evaluated in the χ^2-distribution with 25 degrees of freedom to determine whether the hypothesis is rejected. For the GHSCM we apply the Hilbertian linear model directly with k_n as in Theorem 5.2.1. For each of the 4 models we sample 50, 100, 200, 300, 400 and 500 samples and repeat the experiment 100 times. The results can be seen in Figure 5.3.

For model 1 and 3 where the null is true, we see that both the GCM and the GHSCM appear to hold level. For model 2 and 4 where the null is false, we see that the GCM fails to reject the null, while the GHSCM is able to detect the conditional dependence when n is large enough. While model 2 and 4 were deliberately constructed so that the GCM would fail, we see that the GHSCM allows for detecting more complex dependencies when working with truly infinite-dimensional data.

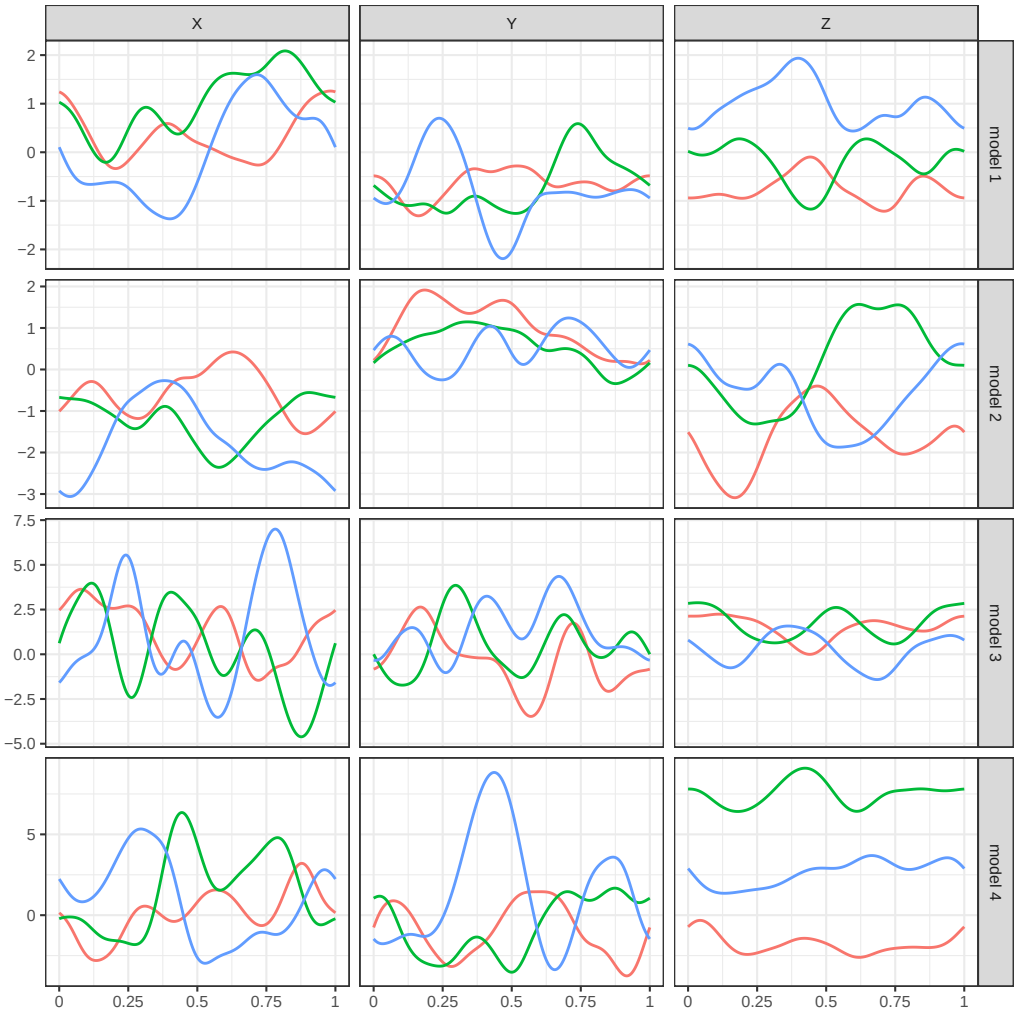


Figure 5.1: Three samples from each of the models (each color marks a sample) viewed as elements of $L^2[0,1]$ using the Fourier basis B_3 described in Example 3.1.8. Note that the scale differs for each of the models.

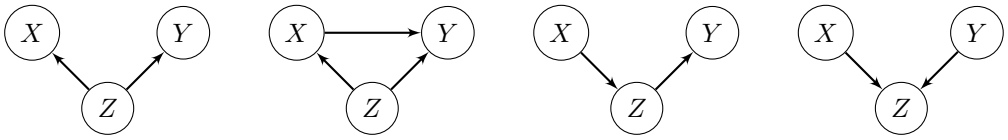


Figure 5.2: Graphs representing each of the four considered models. From left to right they are model 1 through 4.

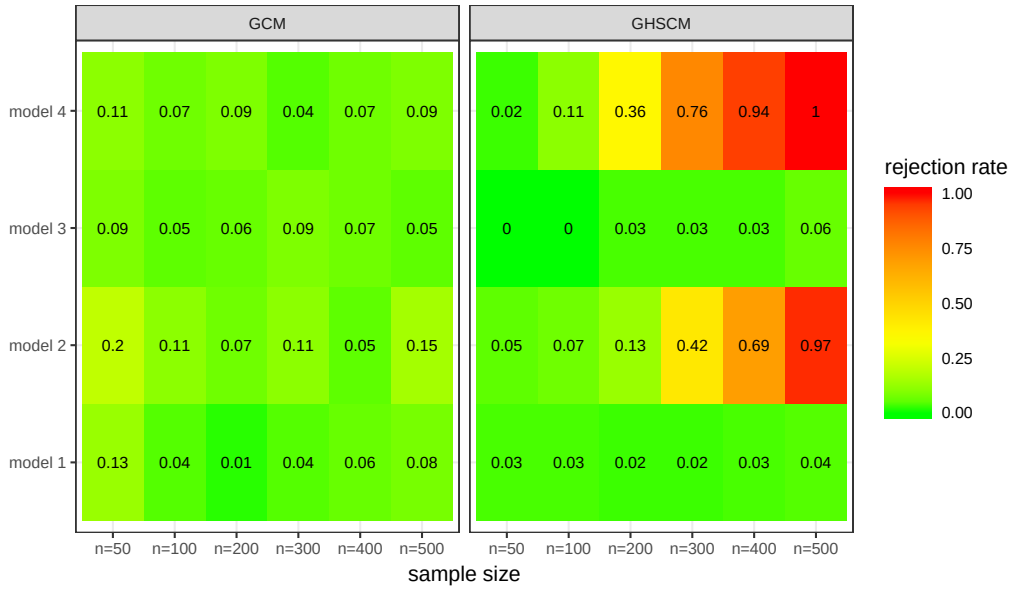


Figure 5.3: Simulation results: While both the GCM and GHSCM hold level in model 1 and 3 only the GHSCM has the ability to detect the conditional dependence in model 2 and 4.

Summary and outlook

In this thesis we constructed a conditional independence test with pointwise asymptotic level for random variables with values in separable Hilbert spaces. We saw that it had applications in the realm of functional data analysis and constructed explicit examples of conditional independence tests for functional data in Theorem 2.3.5 and Theorem 5.1.7. Through a simulation study we also saw that while PCA allows us to use the regular GCM on functional data, we will be able to detect many more types of conditional dependence using the GHSCM.

It would be interesting to extend these results further to Banach space valued random variables. Banach spaces can be significantly more complicated than Hilbert spaces and while non-separable Hilbert spaces are very rare in practice, non-separable Banach spaces occur more often, so generalizations allowing for non-separable spaces would also be interesting. Many of the Hilbert space results in this thesis are motivated by the functional data paradigm that view observed data as functions, typically as elements in the separable Hilbert space $L^2[0,1]$. Results for separable Banach spaces would allow for the functions to be viewed as elements of $C[0,1]$ (the space of continuous real-valued functions on $[0,1]$) and for non-separable Banach spaces would allow for the functions to be viewed as elements of $L^\infty[0,1]$ (the space of (equivalence classes of) essentially bounded real-valued functions on $[0,1]$).

It would also be relevant to see the results of the thesis applied to real data, for instance in causal inference for functional data. There are numerous practical considerations that have not been dealt with in this thesis, such as whether the results still hold when the functional observations are obtained through smoothing of discrete observations. There are also possible computational problems when computing the GHSCM test statistic and bootstrapping the limiting distribution. The aforementioned problems and applications might be more easily solved if one took a different approach to functional data analysis than the one given in this thesis. It is possible to view functional data as stochastic processes rather than Hilbertian random variables and it would be interesting to see the theory of this thesis expressed in that framework.

Appendix

A.1 Measure-theoretic probability theory

In the following, we review some of the fundamental theorems and definitions from measure theory and probability theory, that we will use throughout the thesis. Proofs of the various theorems are omitted for brevity. For a full treatment of the subjects mentioned in this section, we refer to [9], [23] and [27].

A fundamental object in measure theory is the σ-algebra on a set, since this allows us to define a measurable space and then a measure on the space.

Definition A.1.1 (σ-algebras, generators and measurable spaces)**.** Let $\mathcal{X}$ be a set and let $\mathbb{E}$ be a set of subsets of $\mathcal{X}$. We say that $\mathbb{E}$ is a *σ-algebra* if $\mathbb{E}$ satisfies

1. $\mathcal{X} \in \mathbb{E}$,
2. if $A \in \mathbb{E}$, then $A^c \in \mathbb{E}$,
3. if $(A_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{E}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathbb{E}$.

If $\mathbb{E}$ is a σ-algebra, the pair $(\mathcal{X}, \mathbb{E})$ is a *measurable space*.

If $\mathbb{D}$ is some set of subsets of $\mathcal{X}$, we define $\sigma(\mathbb{D})$ to be the smallest σ-algebra on $\mathcal{X}$ containing $\mathbb{D}$ and say that $\mathbb{D}$ *generates* the σ-algebra $\sigma(\mathbb{D})$.

For sets with added structure like topological or metric spaces, we would like to use a σ-algebra that respects the structure. This leads to the definition of the Borel σ-algebra on a space.

Definition A.1.2 (Borel σ-algebra)**.** Let $\mathcal{X}$ be a topological space and let $\mathbb{O}$ be the set of all open subsets of $\mathcal{X}$. We define *the Borel σ-algebra on $\mathcal{X}$* as

$$\mathbb{B}(\mathcal{X}) = \sigma(\mathbb{O}).$$

If $\mathcal{X} = \mathbb{R}$, we simply write $\mathbb{B}$ for the Borel σ-algebra on $\mathbb{R}$.

For concrete spaces we will work exclusively with the Borel σ-algebra. To resolve whether a given class of sets is a generator of a σ-algebra, we often resort to applying Dynkin's lemma.

Theorem A.1.3 (Dynkin's lemma)**.** Let $(\mathcal{X}, \mathbb{E})$ be a measurable space. If $\mathbb{D} \subseteq \mathbb{E}$ satisfies

1. $\mathcal{X} \in \mathbb{D}$,
2. $A, B \in \mathbb{D}$ with $A \subseteq B$ implies $B \backslash B \in \mathbb{D}$,
3. $(A_n)_{n\in\mathbb{N}} \in \mathbb{D}$ with $A_n \subseteq A_{n+1}$ for all n implies $\cup_{n\in\mathbb{N}} A_n \in \mathbb{D}$,

we say that $\mathbb{D}$ is a *Dynkin class*. If $\mathbb{H} \subseteq \mathbb{D}$ is stable under intersections, we have $\sigma(\mathbb{H}) \subseteq \mathbb{D}$.

Once we have a measurable space, we can define a measure on the space.

Definition A.1.4 (Measure and measure space)**.** Let $(\mathcal{X}, \mathbb{E})$ be a measurable space. A function $\mu : \mathbb{E} \to [0, \infty)$ is a *measure*, if

1. $\mu(\varnothing) = 0$,
2. For any sequence of pairwise disjoint sets $(A_n)_{n\in\mathbb{N}}$ in $\mathbb{E}$, $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{i=1}^{\infty} \mu(A_n)$.

The triple $(\mathcal{X}, \mathbb{E}, \mu)$ is a *measure space*.

If $\mu(\mathcal{X}) = 1$, $(\mathcal{X}, \mathbb{E}, \mu)$ is a *probability space* and μ is a *probabiliy measure*.

On $(\mathbb{R}^d, \mathbb{B}^d)$ we define the *d-dimensional Lebesgue measure* m_d as the unique measure assigning sets of the form $[a_1, b_1] \times \cdots \times [a_n, b_n]$ measure $\prod_{i=1}^{n}(b_i - a_i)$.

A crucial part of measure theory is the interplay between mappings and measures. Since measures are defined on σ-algebras these will be the crucial language to understand which mappings are well-behaved and which are not.

Definition A.1.5 (Measurable mappings)**.** Let $(\mathcal{X}, \mathbb{E}$ and $(\mathcal{Y}, \mathbb{D})$ be two measurable spaces and let $f : \mathcal{X} \to \mathcal{Y}$. We say that f *is* $\mathbb{E}$-$\mathbb{D}$-*measurable* if $f^{-1}(D) \in \mathbb{E}$ for all $D \in \mathbb{D}$.

We sometimes simply say that f is measurable when the σ-algebras are obvious.

It is straightforward to show that it is sufficient to check measurability on a generator of $\mathbb{D}$. From this we can see that continuous mappings must be measurable when considering mappings between spaces equipped with Borel σ-algebras.

Having defined measurable mappings, we can define the Lebesgue integral for real-valued functions. Constructing this integral relies on approximating measurable mappings by functions where the integral is straightforward.

Definition A.1.6 (Simple functions and their integral)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $f : \mathcal{X} \to \mathbb{R}$. If f can be written

$$f(x) = \sum_{i=1}^{n} a_i 1_{A_i}(x),$$

for $a_1, \dots a_n \in \mathbb{R}$ and $A_1, \dots A_n \in \mathbb{B}$ are disjoint, we say that f is *simple*. We furthermore define *the Lebesgue integral of the simple function f with respect to μ by*

$$\int f \, \mathrm{d}\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$

Theorem A.1.7 (Approximating Borel-measurable functions by simple functions)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $f : \mathcal{X} \to \mathbb{R}$ be Borel-measurable. Then there exists a sequence of simple functions f_n converging pointwise to f.

Using the above theorem, we can now define the Lebesgue integral.

Definition A.1.8 (Lebesgue integral)**.** Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a measure space and let $f : \mathcal{X} \to \mathbb{R}$ be Borel-measurable and non-negative. We define *the Lebesgue integral of f with respect to μ by*

$$\int f \, \mathrm{d}\mu = \sup \left\{ \int g \, \mathrm{d}\mu \mid f \leqslant g, \ g \text{ is simple} \right\}.$$

For general f, we let $f^+(x) = \max(0, f(x))$ and $f^-(x) = \max(0, -f(x))$ denote the positive and negative parts of f respectively and if both these have finite integrals, we define

$$\int f \, \mathrm{d}\mu = \int f^+ \, \mathrm{d}\mu - \int f^- \, \mathrm{d}\mu.$$

These are the fundamental definitions of Lebesgue integration but to state some of the deeper and often applied theorems, we will turn to random variables since they will be the main way that we will see this theory in the thesis and it is thus convenient to express the results in this language.

Definition A.1.9 (Random variables)**.** A *random variable* $X : \Omega \to \mathcal{X}$ is a measurable mapping from the probability space $(\Omega, \mathbb{F}, P)$ to the measurable space $(\mathcal{X}, \mathbb{E})$. The *distribution* of the random variable is the measure $X(P)$ on $(\mathcal{X}, \mathbb{E})$ and we write $X \sim \nu$ to denote that X has distribution ν.

The vocabulary when working with random variables is sometimes different from when working with general functions but we will use it extensively. In the following we concentrate on real-valued random variables.

Definition A.1.10 (Fundamental operations on real-valued random variables)**.** Let X be a real-valued random variable. The *distribution function* of the random variable X is the function $F : \mathbb{R} \to [0, 1]$ given by

$$F(x) = P(X \leqslant x).$$

The *quantile function* of the random variable X is the generalized inverse of the distribution functions, i.e. it is the function $Q : [0, 1] \to \mathbb{R}$ given by

$$Q(p) = \inf_{x \in \mathbb{R}} \{p \leqslant F(x)\}.$$

The *expectation* $E(X)$ of the random variable X, is the integral of X with respect to P, i.e.

$$E(X) = \int X \, \mathrm{d}P,$$

if $E|X|$ is finite, otherwise it is undefined.

If $E|X|^p$ is finite for some $p > 0$, we say that X *has p-th moment.*

If X has second moment, we define the *variance* of X as

$$\mathrm{Var}(X) = E((X - E(X))^2),$$

and if Y is another real-valued random variable with second moment, we define

$$\mathrm{Cov}(X, Y) = E\left[(X - E(X))(Y - E(Y))\right].$$

These notions can be generalized to multivariate random variables, that take values in $\mathbb{R}^d$.

One of the simplest and yet most often used inequalities, is Markov's inequality:

Theorem A.1.11 (Markov's inequality)**.** Let X be a real-valued random variable. Then for any $\varepsilon > 0$ and $p > 0$, we have

$$P(|X| \geqslant \varepsilon) \leqslant \frac{E|X|^p}{\varepsilon^p}.$$

It is helpful to introduce the notion of independence, to distinguish random variables that affect each other from those that have no effect on each other. Variables that do not affect each other are independent. Independence is defined through σ-algebras as below.

Definition A.1.12 (Independence of σ-algebras)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{F}_1$ and $\mathbb{F}_2$ be sub-σ-algebras of $\mathbb{F}$. If

$$P(F_1 \cap F_2) = P(F_1)P(F_2), \quad \forall F_1 \in \mathbb{F}_1, F_2 \in \mathbb{F}_2,$$

we say that $\mathbb{F}_1$ is *independent* of $\mathbb{F}_2$ and write $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2$.

Definition A.1.13 (Independence of random variables)**.** Let X and Y be random variables defined on the same probability space $(\Omega, \mathbb{F}, P)$ with values in the measurable spaces $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{G})$ respectively. We say that the random variables X and Y are *independent* if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent and we write $X \perp\!\!\!\perp Y$.

We're often interested in sequences of real-valued random variables and how they behave in the limit and to that end, we have a variety of convergence types.

Definition A.1.14 (Convergence of real-valued random variables)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables and X another real-valued random variable.

1. If the set
$$\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$$
has probability 1, we say that X_n *converges almost surely to* X and write $X_n \xrightarrow{a.s.} X$.

2. If for every $\varepsilon > 0$ we have
$$\lim_{n\to\infty} P(|X_n - X| \geqslant \varepsilon) = 0,$$
then we say that X_n *converges in probability to* X and write $X_n \xrightarrow{P} X$.

3. If for every bounded, continuous function $f : \mathbb{R} \to \mathbb{R}$, we have
$$E(f(X_n)) \to E(f(X)) \quad \text{as } n \to \infty,$$
we say that X_n *converges in distribution to* X and write $X_n \xrightarrow{\mathcal{D}} X$.

Some of these notions of convergence imply each other.

Theorem A.1.15 (Relations between modes of convergence)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables and X another real-valued random variable.

1.
$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X$$

2.
$$X_n \xrightarrow{P} X \implies X_n \xrightarrow{\mathcal{D}} X$$

Even though a sequence of integrable random variables converges, it is not obvious if the sequence of expectations converges. The following theorem gives easily checkable and sufficient conditions to ensure convergence of expectations.

Theorem A.1.16 (Dominated Convergence Theorem)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables and X and Z be two other real-valued random variables. Assume that $X_n \xrightarrow{\mathcal{D}} X$, that $|X_n| \leqslant Z$ for all $n \in \mathbb{N}$ and that Z has first moment. Then X has first moment and

$$E(X_n) \to E(X) \quad \text{as } n \to \infty.$$

Convergence in distribution is a very weak form of convergence but we will often use that it is well-behaved when combined with a sequence converging in probability to a constant.

Theorem A.1.17 (Slutsky's theorem)**.** Let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of real-valued random variables and let X be another real-valued random variable such that $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P} c$ for some $c \in \mathbb{R}$. Then

1. $$X_n + Y_n \xrightarrow{\mathcal{D}} X + c$$
2. $$X_n Y_n \xrightarrow{\mathcal{D}} cX$$
3. $$\frac{X_n}{Y_n} \xrightarrow{\mathcal{D}} \frac{X}{c}, \quad \text{if } c \neq 0$$

Sequences that are independent and identically distributed (i.i.d) are particularly well-behaved as the following deep and often-applied results show.

Theorem A.1.18 (Law of Large Numbers (LLN))**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed, real-valued random variables. If $E|X_1| < \infty$ then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} E(X_1)$$

Theorem A.1.19 (Central Limit Theorem (CLT))**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed, real-valued random variables. If $E|X_1|^2 < \infty$ then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E(X_1)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{Var}(X_1))$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean $\mu \in \mathbb{R}$ and variance σ^2 where $\sigma > 0$.

Much work has been dedicated to generalizing the CLT to settings where the distributions of the variables in question are not necessarily the same. This lead to the central limit theorems for triangular arrays.

Theorem A.1.20 (Lyapounov's central limit theorem)**.** Let $(X_{nk})_{1\leqslant k\leqslant n}$ be a triangular array of real-valued random variables. Set $S_n = \sum_{k=1}^{n} X_{nk}$. Assume that

1. For all $n \geqslant 1$, the family $(X_{nk})_{k\leqslant n}$ is independent.
2. $E(X_{nk}) = 0$ for all $1 \leqslant k \leqslant n$.
3. $\mathrm{Var}(S_n) \to 1$ as $n \to \infty$.
4. There exists $\eta > 0$ so that
$$\lim_{n\to\infty} \sum_{i=1}^{n} E|X_{ni}|^{2+\eta} = 0.$$

Then $S_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

Sometimes we would not only like to approximate a single expectation but a whole distribution simultaneously. For real-valued random variables this can be done using the empirical distribution function, which the following theorem shows, approximates the true distribution function in the i.i.d setting.

Theorem A.1.21 (Glivenko-Cantelli theorem)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed, real-valued random variables with common distribution function $F(x)$. Define the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(X_i \geqslant x)}(x).$$

Then

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

This is the theoretical basis for the validity of bootstrapping.

We will also apply the theory of conditional expectations. The expectation represents our "best guess" about the value of random variable when no information is given. The conditional expectation given a σ-algebra represents our best guess, given the information contained in the σ-algebra.

Definition A.1.22 (Conditional expectation)**.** Let X be a random variable defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in $(\mathbb{R}, \mathbb{B})$. Assume further that X is integrable, i.e. $E|X| < \infty$ and let $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. Then *the conditional expectation of X given* $\mathbb{D}$ is denoted by $E(X \mid \mathbb{D})$ and is the almost surely unique random variable satisfying

1. $E(X \mid \mathbb{D})$ is $\mathbb{D}$-measurable,
2. for every $D \in \mathbb{D}$
$$\int_D E(X \mid \mathbb{D}) \, \mathrm{d}P = \int_D X \, \mathrm{d}P.$$

If Y is another random variable on the same probability space with values in some measurable space, we define $E(X \mid Y)$ to mean $E(X \mid \sigma(Y))$.

If we write $E(X \mid Y, Z)$ for a third random variable Z, we mean $E(X \mid \sigma(Y, Z))$ and similarly if $\mathbb{H}$ is another sub-σ-algebra of $\mathbb{F}$ we write $E(X \mid \mathbb{D}, \mathbb{H})$ to mean $E(X \mid \sigma(\mathbb{D}, \mathbb{H}))$.

It is worthwhile to note that if X is real-valued and measurable wrt. to the σ-algebra generated by a random variable Y on an arbitrary measure space, this implies that X can be written as the composition of Y with a measurable function. This is the content of the Doob-Dynkin lemma.

Theorem A.1.23 (Doob-Dynkin lemma)**.** Let X be a real-valued random variable defined on $(\Omega, \mathbb{F}, P)$ and let Y be another random variable on the same probability space with values in the measure space $(\mathcal{Y}, \mathbb{E})$. Then X is $\sigma(Y)$-measurable if and only if there exists a $\mathbb{E} - \mathbb{B}$-measurable function $\phi : \mathcal{Y} \to \mathbb{R}$ so that

$$X = \phi \circ Y.$$

This lets us define what is meant by a conditional expectation given that $Y = y$.

Definition A.1.24 (Conditional expectation givne value of variable)**.** Let X be a real-valued random variable defined on $(\Omega, \mathbb{F}, P)$ and let Y be another random variable on the same probability space with values in the measure space $(\mathcal{Y}, \mathbb{E})$. Assume that X has finite first moment, so that $E(X \mid Y)$ exists. By the Doob-Dynkin lemma there exists a measurable $\phi : \mathcal{Y} \to \mathbb{R}$ such that $E(X \mid Y) = \phi \circ Y$. We define $E(X \mid Y = y) := \phi(y)$ and call this *the conditional expectation of X given $Y = y$.*

Conditional expectations have several nice properties.

Theorem A.1.25 (Properties of conditional expectation)**.** Let X be a random variable defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in $(\mathbb{R}, \mathbb{B})$. Assume further that X is integrable, i.e. $E|X| < \infty$ and let $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. Then

1. If $\mathbb{H} \subseteq \mathbb{D}$ is a third σ-algebra, we have

$$E(X \mid \mathbb{H}) = E(E(X \mid \mathbb{H}) \mid \mathbb{D}) = E(E(X \mid \mathbb{D}) \mid \mathbb{H}),$$

2. If $\sigma(X)$ and $\mathbb{D}$ are independent then

$$E(X \mid \mathbb{D}) = E(X),$$

3. If X is $\mathbb{D}$-measurable then

$$E(X \mid \mathbb{D}) = X,$$

4. If Y is another real-valued integrable random variable, XY is integrable and X is $\mathbb{D}$-measurable, we have

$$E(XY \mid \mathbb{D}) = XE(Y \mid \mathbb{D}).$$

We also have a version of dominated convergence for conditional expectations.

Theorem A.1.26 (Conditional dominated convergence theorem)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables defined on the probability space $(\Omega, \mathbb{F}, P)$ and let X be another such random variable. Assume further that for all n, X_n is integrable, X is integrable and let $\mathbb{D}$ be a sub-σ-algebra of $\mathbb{F}$. Assume that $X_n \overset{a.s.}{\to} X$ and $|X_n| \leqslant Y$ for all n for some integrable real-valued random variable Y. Then

$$E(X_n \mid \mathbb{D}) \overset{a.s.}{\to} E(X \mid \mathbb{D}).$$

A.2 Established results and definitions from analysis and linear algebra

In the following we review some of the fundamental theorems and definitions from analysis and linear algebra, that are used throughout the thesis. Proofs are omitted for brevity. For a full treatment, see [23], [12] and [10].

The fundamental object of study in linear algebra is the vector space.

Definition A.2.1 (Vector space)**.** Let F be a field and let V be a set. Let $+ : V \times V \to V$ and $\cdot : F \times V \to V$ be mappings. We say that V *is a vector space over* F if

1. $\forall x, y, z \in V : (x + y) + z = x + (y + z)$
2. $\exists 0 \in V \ \forall x \in V : x + 0 = 0 + x = x$
3. $\forall x \in V \ \exists y \in V : x + y = y + x = 0$

4. $\forall x, y \in V : x + y = y + x$
5. $\forall x \in V \ \forall a, b \in F : a \cdot (b \cdot V) = (ab) \cdot V$
6. $\forall x, y \in V \ \forall a \in F : a \cdot (x + y) = a \cdot x + a \cdot y$
7. $\forall x \in V \ \forall a, b \in F : (a + b) \cdot x = a \cdot x + b \cdot x$
8. $\forall x \in V : 1 \cdot x = x$

The canonical examples of vector spaces are $\mathbb{R}^d$ over $\mathbb{R}$ with the usual addition and multiplication. In this thesis we will solely consider vector spaces over $\mathbb{R}$. Some vector spaces have added structure such as the notion of a length (a norm) or a notion of orthogonality (an inner product).

Definition A.2.2 (Normed space)**.** Let V be a vector space over F and let $\|\cdot\| : V \to F$ be a mapping. We say that V *is a normed space* and that $\|\cdot\|$ *is a norm on* V if

1. $\|x\| = 0 \iff x = 0$
2. $\forall x \in V \ \forall a \in F : \|ax\| = a\|x\|$
3. $\forall x, y \in V : \|x + y\| \leqslant \|x\| + \|y\|$

Definition A.2.3 (Inner product space)**.** Let V be a vector space over F and let $\langle \cdot, \cdot \rangle : V \times V \to F$ be a mapping. We say that V *is an inner product space* and that $\langle \cdot, \cdot \rangle$ *is an inner product on* V if

1. $\forall x \in V : \langle x, x \rangle \geqslant 0$
2. $\forall x, y \in V \ \forall a \in F : \langle a \cdot x, y \rangle = a \langle x, y \rangle$
3. $\forall x, y, z \in V : \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
4. $\forall x, y \in V : \langle x, y \rangle = \overline{\langle y, x \rangle}$

where $\overline{a}$ denotes the conjugate of a.

Every inner product space is also a normed space, since setting $\|x\| = \sqrt{\langle x, x \rangle}$ becomes a norm. In inner product spaces we have the crucial Cauchy-Schwarz inequality.

Theorem A.2.4 (Cauchy-Schwarz inequality)**.** Let V be an inner product space over F with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\|\cdot\|$. Then for all $x, y \in V$, we have

$$\langle x, y \rangle^2 \leqslant \|x\|^2 \|y\|^2.$$

While we almost solely consider vector spaces in this thesis, there are some more general concepts that are relevant to consider, in particular those related to limits and distances on arbitrary spaces.

Definition A.2.5 (Topology and topological spaces)**.** Let $\mathcal{X}$ be some set. We say that a collection of subsets τ is a *topology* if

1. $\varnothing \in \tau$ and $\mathcal{X} \in \tau$
2. The union (countable or uncountable) of arbitrary elements of τ is again in τ
3. Finite intersections of elements in τ are again in τ

$(\mathcal{X}, \tau)$ is called a *topological space* and the elements of τ are the *open sets* of $\mathcal{X}$. If A is a subset of X, then A is said to be *closed* if $A^c \in \tau$. For an arbitrary subset of X, A, we define *the closure of* A as the smallest closed set containing A and denote it $\bar{A}$.

A topology is the fundamental tool for investigating convergence and continuity. When we're working with uncountably infinite spaces, we would like to reduce our problems to countable problems. Broadly speaking this is possible in separable spaces.

Definition A.2.6 (Dense sets and separability)**.** Let $(\mathcal{X}, \tau)$ be a topological space. A set $A \subseteq \mathcal{X}$ is said to be *dense* if $\bar{A} = \mathcal{X}$. (X, τ) is said to be *separable* if it contains a countable, dense subset.

While topologies are fundamental, we will often work with them indirectly by working with a metric, that induces a topology.

Definition A.2.7 (Metric space)**.** Let $\mathcal{X}$ be a set and let $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a mapping. We say that M *is a metric space* and d *is a metric* if

1. $d(x, y) = 0 \iff x = y$
2. $\forall x, y \in \mathcal{X} : d(x, y) = d(y, x)$
3. $\forall x, y, z \in \mathcal{X} : d(x, z) \leqslant d(x, y) + d(y, z)$

A subset A of $\mathcal{X}$ is open in the metric space, if for every $a \in A$, there exists $\varepsilon > 0$ so that the ε-ball around a is enclosed in A, i.e.

$$B(a, \varepsilon) = \{x \in M \mid d(a, x) < \varepsilon\} \subseteq A.$$

The collection of all open sets using the metric forms a topology. One should also note that any normed vector space is also a metric space by setting $d(x,y) = \|x-y\|$. Metric spaces give us a notion of distance on a space. We would like for the spaces we consider to have no holes, which leads to the following definition.

Definition A.2.8 (Sequences and completeness)**.** Let $\mathcal{X}$ be a space with metric d and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{X}$. We say that x_n *converges to* x if

$$\lim_{n\to\infty} d(x_n, x) = 0.$$

We say that x_n *is a Cauchy sequence* if

$$\lim_{n,m\to\infty} d(x_n, x_m) = 0.$$

We say that $\mathcal{X}$ is *complete* if every Cauchy sequence converges to some $x \in \mathcal{X}$.

A.3 Auxiliary results

In this section we prove results from the main thesis, that were deemed to be too long in relation to their importance.

Theorem A.3.1 (Equivalent definition of conditional independence)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{F}_1$, $\mathbb{F}_2$ and $\mathbb{F}_3$ be sub-σ-algebras of $\mathbb{F}$. $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3$ if and only if

$$P(1_{F_1} \mid \mathbb{F}_2, \mathbb{F}_3) = P(1_{F_1} \mid \mathbb{F}_3), \tag{$*$}$$

for all $F_1 \in \mathbb{F}_1$.

Proof.
We follow the same strategy as the proof given in [5] Proposition 2.3.28.

Assuming that $(*)$ holds, we get conditional independence straight away by properties of conditional expectation, since for any $F_1 \in \mathbb{F}_1$ and $F_2 \in \mathbb{F}_2$, we have

$$\begin{aligned} E(1_{F_1}1_{F_2} \mid \mathbb{F}_3) &= E(E(1_{F_1}1_{F_2} \mid \mathbb{F}_2, \mathbb{F}_3) \mid \mathbb{F}_3) = E(E(1_{F_1} \mid \mathbb{F}_2, \mathbb{F}_3)1_{F_2} \mid \mathbb{F}_3) \\ &= E(E(1_{F_1} \mid \mathbb{F}_3)1_{F_2} \mid \mathbb{F}_3) = E(1_{F_2} \mid \mathbb{F}_3)E(1_{F_1} \mid \mathbb{F}_3). \end{aligned}$$

Assuming instead conditional independence, we will prove $(*)$ using a Dynkin class argument. Letting $F_1 \in \mathbb{F}_1$ be given, note first that the set

$$\mathbb{H} = \{F_2 \cap F_3 \mid F_2 \in \mathbb{F}_2, F_3 \in \mathbb{F}_3\}$$

generates $\sigma(\mathbb{F}_2, \mathbb{F}_3)$. This holds since it is straightforward to see that any set in $\mathbb{F}_2 \cup \mathbb{F}_3$ must also be in $\mathbb{H}$, so $\sigma(\mathbb{F}_2, \mathbb{F}_3) \subseteq \mathbb{H}$. It is also straightforward to see that $\mathbb{H} \subseteq \sigma(\mathbb{F}_2, \mathbb{F}_3)$. $\mathbb{H}$ is also stable under intersections, so Dynkin's lemma yields that $\sigma(\mathbb{H}) = \sigma(\mathbb{F}_2, \mathbb{F}_3)$.

Returning to the problem, we need to show that $E(1_{F_1} \mid \mathbb{F}_2, \mathbb{F}_3) = E(1_{F_1} \mid \mathbb{F}_3)$ and will do so straight from the definition. Obviously $E(1_{F_1} \mid \mathbb{F}_3)$ is $\sigma(\mathbb{F}_2, \mathbb{F}_3)$-measurable, so we only need to show that

$$\int_D 1_{F_1} \, \mathrm{d}P = \int_D E(1_{F_1} \mid \mathbb{F}_3) \, \mathrm{d}P \tag{$\dagger$}$$

for all $D \in \sigma(\mathbb{F}_2, \mathbb{F}_3)$. To that end define $\mathbb{D}_{F_1}$ to be the collection of all sets in $\sigma(\mathbb{F}_2, \mathbb{F}_3)$ satisfying ($\dagger$). If we can show that $\mathbb{H} \subseteq \mathbb{D}_{F_1}$ and that $\mathbb{D}_{F_1}$ is a Dynkin class, we will be done by Dynkin's lemma, since then $\sigma(\mathbb{F}_2, \mathbb{F}_3) = \sigma(\mathbb{H}) \subseteq \mathbb{D}_{F_1} \subseteq \sigma(\mathbb{F}_2, \mathbb{F}_3)$. To show that $\mathbb{H} \subseteq \mathbb{D}_{F_1}$, take some $H \in \mathbb{H}$, i.e. $H = F_2 \cap F_3$ for some $F_2 \in \mathbb{F}_2$ and $F_3 \in \mathbb{F}_3$ and note that

$$\begin{aligned}
&\int_H 1_{F_1} \, \mathrm{d}P = \int_{F_3} 1_{F_1} 1_{F_2} \, \mathrm{d}P = \int_{F_3} E(1_{F_1} 1_{F_2} \mid \mathbb{F}_3) \, \mathrm{d}P = \int_{F_3} E(1_{F_1} \mid \mathbb{F}_3) E(1_{F_2} \mid \mathbb{F}_3) \, \mathrm{d}P \\
&= \int_{F_3} E\big[E(1_{F_1} \mid \mathbb{F}_3) 1_{F_2} \mid \mathbb{F}_3\big] \, \mathrm{d}P = \int_{F_3} E(1_{F_1} \mid \mathbb{F}_3) 1_{F_2} \, \mathrm{d}P = \int_H E(1_{F_1} \mid \mathbb{F}_3) \, \mathrm{d}P,
\end{aligned}$$

by various properties of the conditional expectation.

To show that $\mathbb{D}_{F_1}$ is a Dynkin class, we note first that by the tower property $\Omega \in \mathbb{D}_{F_1}$. To show that $\mathbb{D}_{F_1}$ is stable under set difference, we see that for $D_1, D_2 \in \mathbb{D}_{F_1}$ with $D_1 \subseteq D_2$, we have

$$\begin{aligned}
\int_{D_2 \setminus D_1} 1_{F_1} \, \mathrm{d}P &= \int 1_{F_1} 1_{D_2} - 1_{F_1} 1_{D_1} \, \mathrm{d}P = \int_{D_2} 1_{F_1} \, \mathrm{d}P - \int_{D_1} 1_{F_1} \, \mathrm{d}P \\
&= \int_{D_2} E(1_{F_1} \mid \mathbb{F}_3) \, \mathrm{d}P - \int_{D_1} E(1_{F_1} \mid \mathbb{F}_3) \, \mathrm{d}P \\
&= \int (1_{D_2} - 1_{D_1}) E(1_{F_1} \mid \mathbb{F}_3) \, \mathrm{d}P = \int_{D_2 \setminus D_1} E(1_{F_1} \mid \mathbb{F}_3) \, \mathrm{d}P,
\end{aligned}$$

by using that $1_{D_2 \setminus D_1} = 1_{D_2} - 1_{D_1}$ and the fact that both D_1 and D_2 are in $\mathbb{D}_{F_2}$. Finally taking an increasing sequence $(D_n)_{n \in \mathbb{N}}$ in $\mathbb{D}_{F_1}$, we have

$$\begin{aligned}
\int_{\cup_{n \in \mathbb{N}} D_n} 1_{F_1} \, \mathrm{d}P &= \int 1_{\cup_{n \in \mathbb{N}} D_n} 1_{F_1} \, \mathrm{d}P = \int \lim_{n \to \infty} 1_{D_n} 1_{F_1} \, \mathrm{d}P = \lim_{n \to \infty} \int_{D_n} 1_{F_1} \, \mathrm{d}P \\
&= \lim_{n \to \infty} \int_{D_n} E(1_{F_1} \mid \mathbb{F}_3) \, \mathrm{d}P = \int_{\cup_{n \in \mathbb{N}} D_n} E(1_{F_1} \mid \mathbb{F}_3) \, \mathrm{d}P,
\end{aligned}$$

by dominated convergence. □

Theorem A.3.2 (Alternative characterization of conditional independence)**.** Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{F}_1$, $\mathbb{F}_2$ and $\mathbb{F}_3$ be sub-σ-algebras of $\mathbb{F}$. $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3$ if and only if

$$(\mathbb{F}_1, \mathbb{F}_3) \perp\!\!\!\perp (\mathbb{F}_2, \mathbb{F}_3) \mid \mathbb{F}_3.$$

Proof.
We will show this using the properties of conditional independence derived in Theorem 2.1.9.

It is straightforward by applying decomposition and symmetry that $(\mathbb{F}_1, \mathbb{F}_3) \perp\!\!\!\perp (\mathbb{F}_2, \mathbb{F}_3) \mid \mathbb{F}_3 \implies \mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3$.

To show the converse, note first that we have $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_3 \mid (\mathbb{F}_2, \mathbb{F}_3)$ trivially since for $F_1 \in \mathbb{F}_1$ and $F_3 \in \mathbb{F}_3$

$$E(1_{F_1} 1_{F_3} \mid \mathbb{F}_2, \mathbb{F}_3) = 1_{F_3} E(1_{F_1} \mid \mathbb{F}_2, \mathbb{F}_3) = E(1_{F_3} \mid \mathbb{F}_2, \mathbb{F}_3) E(1_{F_1} \mid \mathbb{F}_2, \mathbb{F}_3).$$

Thus since both $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_2 \mid \mathbb{F}_3$ and $\mathbb{F}_1 \perp\!\!\!\perp \mathbb{F}_3 \mid (\mathbb{F}_2, \mathbb{F}_3)$, we can apply contraction to get $\mathbb{F}_1 \perp\!\!\!\perp (\mathbb{F}_2, \mathbb{F}_3) \mid \mathbb{F}_3$. Using symmetry and noting that $(\mathbb{F}_2, \mathbb{F}_3) \perp\!\!\!\perp \mathbb{F}_3 \mid (\mathbb{F}_1, \mathbb{F}_3)$ by similar arguments as earlier, we can again use contraction to get the desired result. □

Bibliography

[1] Viktor Bengs and Hajo Holzmann. *Uniform approximation in classical weak convergence theory*. 2019. eprint: arXiv:1903.09864.

[2] Patrick Billingsley, ed. *Convergence of Probability Measures*. John Wiley & Sons, Inc., July 1999. DOI: 10.1002/9780470316962. URL: https://doi.org/10.1002/9780470316962.

[3] Denis Bosq. *Linear Processes in Function Spaces*. Springer New York, 2000. DOI: 10.1007/978-1-4612-1154-9. URL: https://doi.org/10.1007/978-1-4612-1154-9.

[4] T. Tony Cai and Peter Hall. "Prediction in functional linear regression". In: *Ann. Statist.* 34.5 (Oct. 2006), pp. 2159–2179. DOI: 10.1214/009053606000000830. URL: https://doi.org/10.1214/009053606000000830.

[5] Panayiota Constantinou. "Conditional Independence and Applications in Statistical Causality". PhD thesis. University of Cambridge, Girton College, May 2013.

[6] Christophe Crambes and André Mas. "Asymptotics of prediction in functional linear regression with functional outputs". In: *Bernoulli* 19.5B (Nov. 2013), pp. 2627–2651. DOI: 10.3150/12-bej469. URL: https://doi.org/10.3150/12-bej469.

[7] J. J. DAUDIN. "Partial association measures and an application to qualitative regression". In: *Biometrika* 67.3 (1980), pp. 581–590. DOI: 10.1093/biomet/67.3.581. URL: https://doi.org/10.1093/biomet/67.3.581.

[8] Ernst Hansen. *Introduktion til Matematisk Statistik*. University of Copenhagen, 2012.

[9] Ernst Hansen. *Measure theory*. University of Copenhagen, 2006.

[10] Lars Hesselholt and Nathalie Wahl. *Lineær Algebra*. University of Copenhagen, 2016.

[11] E. Hille and R.S. Phillips. *Functional Analysis and Semi Groups*. Colloquium Publications - American Mathematical Society. American Mathematical Soc., 1957. URL: https://books.google.dk/books?id=hPpQAAAAMAAJ.

[12] Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, Ltd, May 2015. DOI: 10.1002/9781118762547. URL: https://doi.org/10.1002/9781118762547.

[13] Jens Ledet Jensen and Michael Sørensen. *Statistical Principles: A First Course*. University of Copenhagen, 2017.

[14] Maximilian Kasy. "Uniformity and the Delta Method". In: *Journal of Econometric Methods* 8.1 (Jan. 2019), pp. 1–19. URL: https://ideas.repec.org/a/bpj/jecome/v8y2019i1p19n9.html.

[15] S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996. ISBN: 9780191591228. URL: https://books.google.dk/books?id=mGQWkx4guhAC.

[16] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer Berlin Heidelberg, 1991. DOI: 10.1007/978-3-642-20212-4. URL: https://doi.org/10.1007/978-3-642-20212-4.

[17] E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer New York, 2005. DOI: 10.1007/0-387-27605-x. URL: https://doi.org/10.1007/0-387-27605-x.

[18] Jan van Neerven. *Stochastic Evolution Equations*. 2008. URL: https://fa.its.tudelft.nl/~neerven/publications/notes/ISEM.pdf.

[19] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[20] Philip T. Reiss et al. "Methods for Scalar-on-Function Regression". In: *International Statistical Review* 85.2 (Feb. 2016), pp. 228–249. DOI: 10.1111/insr.12163. URL: https://doi.org/10.1111/insr.12163.

[21] Joseph P. Romano and Azeem M. Shaikh. "On the uniform asymptotic validity of subsampling and the bootstrap". In: *Ann. Statist.* 40.6 (Dec. 2012), pp. 2798–2822. DOI: 10.1214/12-AOS1051. URL: https://doi.org/10.1214/12-AOS1051.

[22] Frank S. Scalora. "Abstract martingale convergence theorems." In: *Pacific J. Math.* 11.1 (1961), pp. 347–374. URL: https://projecteuclid.org:443/euclid.pjm/1103037558.

[23] René L Schilling. *Measures, Integrals and Martingales*. Cambridge University Press, 2017.

[24] Štefan Schwabik and Ye Guoju. *Topics in Banach Space Integration*. WORLD SCIENTIFIC, Aug. 2005. DOI: 10.1142/5905. URL: https://doi.org/10.1142/5905.

[25] Rajen D. Shah and Jonas Peters. "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure". In: *arXiv preprint arXiv:1804.07203* (2018).

[26] Hyejin Shin and Myung Hee Lee. "On prediction rate in partial functional linear regression". In: *Journal of Multivariate Analysis* 103.1 (Jan. 2012), pp. 93–106. DOI: 10.1016/j.jmva.2011.06.011. URL: https://doi.org/10.1016/j.jmva.2011.06.011.

[27] A. Sokol and A. Rønn-Nielsen. *Advanced Probability*. Department of Mathematical Sciences, University of Copenhagen, 2014. ISBN: 9788770789738. URL: https://books.google.dk/books?id=uqQArgEACAAJ.

[28] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability Distributions on Banach Spaces*. Springer Netherlands, 1987. DOI: 10.1007/978-94-009-3873-1. URL: https://doi.org/10.1007/978-94-009-3873-1.

[29] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. "Functional linear regression analysis for longitudinal data". In: *Ann. Statist.* 33.6 (Dec. 2005), pp. 2873–2903. DOI: 10.1214/009053605000000660. URL: https://doi.org/10.1214/009053605000000660.