UNIVERSITY OF CALIFORNIA
SANTA CRUZ

REGRESSION THROUGH FUNCTIONAL DATA ANALYSIS

A thesis submitted in partial satisfaction
of the requirements for the degree of

MASTER OF ARTS

in

MATHEMATICS

by

Kevin Amilcar Guillen

September 2023

The Thesis of Kevin Amilcar Guillen
is approved:

Professor Pedro Morales-Almazan, Chair

Professor François Monard

Professor Torsten Ehrhardt

Peter Biehl
Vice Provost and Dean of Graduate Studies

Contents

Abstract

REGRESSION THROUGH FUNCTIONAL DATA ANALYSIS

by

Kevin Amilcar Guillen

This thesis delves into the world of Functional Data Analysis (FDA) and its analog of Principal Component Analysis (PCA) termed Functional PCA (FPCA). While a brief primer on traditional PCA sets the stage, the main emphasis is on the richness of FDA—a branch of statistics focusing on data represented as curves or surfaces—and the nuances that distinguish it from conventional data analysis techniques. From this foundation, the thesis elaborates on FPCA and its inherent capability to handle the infinite-dimensional nature of functional data, with methodologies rooted in Hilbert spaces. The core exploration revolves around extensions of standard definitions and theorems in statistics, then ends with an application of FPCA in the realm of forecasting. Empirical demonstrations highlight the potential and advantages of utilizing FPCA for prediction tasks. In synthesizing the areas of traditional statistics and functional analysis, this thesis highlights FPCA in the landscape of data analysis.

Dedicated to my parents, Kathy and Tomas Guillen.

Acknowledgments

I am grateful to my advisor, Professor Pedro Morales-Almazan, for taking me on and for his kind guidance throughout my thesis journey. A special thanks to Professor François Monard for suggesting the idea to write a thesis and sharing his awesome knowledge of complex analysis, and to Professor Torsten Ehrhardt for his excellent instruction in functional analysis. Thanks again to both for being on my reading committee. I thank them all once more, as well as many of the other professors in the math department, for being great instructors during my time as both a graduate and undergraduate student at UCSC.

I owe a thanks to Professor David Draper for deepening my interest in probability and statistics.

I thank my parents for their unwavering support and faith in me, and for sharing every bit of love and wisdom they possess. Thanks to my sister, Andrea, for all of our conversations, always looking out for me, and being a true friend. Love and thanks to my extended family for their love and support, and to my friends Ethan, Gurjot, and Thomas.

Furthermore, I want to thank my partner, Tú Anh, for all her support and patience. She's been my best friend, always lifting my spirits, and consistently giving me something to look forward to. I extend my gratitude to her parents and sister as well for supporting me as a son and brother.

Lastly, my thanks to those of us in the basement, specifically Brian, David, Deewang, Jennifer, Malachi, and Ryan. Thanks for all the laughter, conversations, and guidance.

List of Figures

Chapter 1

Introduction

1.1 Introduction

Functional data analysis (FDA) is a statistical branch analyzing data represented as curves or surfaces, termed functional data. Originating early as the 1950s, it matured in the late 1980s due to Ramsay and Dalzell [8]. This data is inherently infinite dimensional, and samples of functional data are considered to be functions, so almost all analysis is performed in Hilbert and Sobolev spaces [2]. Due to this change in setting from traditional statistical analysis, much work has to be put in to extend traditional methods. Our focus will be on the extension of one of these methods, Principal Component Analysis (PCA) for forecasting. Since FDA is a blend of statistics and functional analysis we must first have a general understanding of both areas.

1.2 Statistics

There are a variety of philosophies, methodologies, and tools for conducting statistical analysis. Here we are going to briefly carve through one important method

of analysis, since the functional analog of this method is our focus.

Definition 1.2.1. *Let X be a random variable with a finite set of possible outcomes $x_1, \ldots, x_k$ each with probability $p_1, \ldots, p_k$ respectively. The expectation of X is defined as,*

$$\mathbb{E}[X] = x_1 p_1 + \cdots + x_k p_k,$$

Take X to be a p-dimensional random vector $(X_1, \ldots, X_p)^T$ having (co-)variance matrix ,

$$K = \mathbb{E}[(X - m)(X - m)^T]$$

where m is the expectation of X. The (co-)variance matrix has the following eigenvalue-eigenvector decomposition,

$$K = \sum_{j=1}^{p} \lambda_j e_j e_j^T,$$

for eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and their associated orthonormal eigenvectors $e_j = (e_1, \ldots, e_{pj})^T$, $j = 1, \ldots, p$ that satisfy,

$$e_i^T K e_j = \lambda_j \delta_{ij},$$

where δ_{ij} is the Kronecker delta. Using the eigenvector decomposition we can define principal components,

$$Z_j = e_j^T (X - m),$$

which are linear combinations of original variables with the weight of e_{ij} applied to X_i in the jth component, representing its importance to Z_j, measured specifically by,

$$Cov(Z_j, X_j) = \lambda_j e_{ij}.$$

Using this, one can represent X through,

$$X = m + \sum_{j=1}^{p} Z_j e_j,$$

since $e_1, \ldots, e_p$ provide an orthonormal basis for $\mathbb{R}^p$. In language, this is just showing that we can represent X through a weighted sum of eigenvectors, which are obtained from K, with the weights being the eigenvalues (and they are uncorrelated). Under this perspective the eigenvalues are referred to as scores, since their values represent how much variance is captured.

During analysis, one usually chooses to use n of these components where $n < p$, since using n components measures the relationship between the variables in X up to an arbitrary percentage. So, one does lose information from using fewer components, but dimensional reduction is achieved. The amount of information captured or lost as a consequence of using n components is easily measured since the total variance of X is,

$$V = \sum_{j=1}^{p} \lambda_j,$$

where the variance of the jth component is,

$$Var(Z_j) = e_j^T K e_j = \lambda_j,$$

so, using a subset of components $\{Z_{j_k}\}_{k \in I}$ (where I is some index), we retain,

$$\frac{1}{V} \sum_{k \in I} \lambda_{j_k},$$

of the information. Since one knows how much information each component captures, one can leverage them to perform outlier detection, pattern detection, forecasting, and many other applications. This is the basis of principal component analysis.

Chapter 2

Functional Analysis

Before getting into functional data analysis we must first lay out key definitions and theorems that will be needed. All of these definitions and theorems are common, and their proofs can be found in [1], [3], [11] or elsewhere.

2.1 Preliminaries

Definition 2.1.1. *Let $(E, \mathscr{B}, \mu)$ be a measure space and for $p \in [1, \infty)$, denote by $L^p(E, \mathscr{B}, \mu)$ the collection of measurable functions f on E that satisfy $\int_E |f|^p \, d\mu < \infty$. Define,*

$$\|f\|_p = \left(\int_E |f|^p \, d\mu \right)^{1/p} \tag{2.1}$$

when $f \in L^p(E, \mathcal{B}, \mu)$.

Definition 2.1.2. *A function $f : E \to \mathbb{X}$ is called simple if it can be represented as*

$$f(\omega) = \sum_{i=1}^{k} \mathbb{I}_{E_i}(\omega) g_i$$

for some finite k, $g_i \in \mathbb{X}$, and $E_i \in \mathscr{B}$.

Definition 2.1.3. *Any simple function* $f(\omega) = \sum_{i=1}^{k} \mathbb{I}_{E_i}(\omega) g_i$ *with* $\mu(E_i) < \infty$ *for all* i *is said to be integrable and its Bochner integral is defined as*

$$\int_E f d\mu = \sum_{i=1}^{k} \mu(E_i) g_i.$$

Definition 2.1.4. *A measurable function* f *is said to be Bochner integrable if there exists a sequence* $\{f_n\}$ *of simple and Bochner integrable functions such that*

$$\lim_{n\to\infty} \int_E \|f_n - f\|_p \, d\mu = 0.$$

In this case, the Bochner integral of f *is defined as*

$$\int_E f d\mu = \lim_{n\to\infty} \int_E f_n d\mu.$$

2.2 Linear Functionals and Operators

From here forward, $\boldsymbol{\mathcal{B}}(\mathbb{H}, \mathbb{R})$ or $\boldsymbol{\mathcal{B}}(\mathbb{H})$ refers to the dual space of $\mathbb{H}$, Hilber-Schmidt operators will be abbreviated as HS, and the collection of HS operators in the dual of $\mathbb{H}$ is denoted by $\boldsymbol{\mathcal{B}}_{HS}(\mathbb{H}, \mathbb{R})$, and the norm in the Hilbert space this collection makes is denoted as $\|\cdot\|_{HS}$.

Theorem 2.2.1. *Let* X_1, X_2 *be Banach spaces,* f *a Bochner integrable function from* E *to* X_1*, and* $\mathcal{F} \in \boldsymbol{\mathcal{B}}(X_1, X_2)$*. Then* $\mathcal{F}f$ *is Bochner integrable and,*

$$\mathcal{F}\left(\int_E f d\mu\right) = \int_E \mathcal{F} f d\mu.$$

Corollary 2.2.2. *Let* $\mathbb{H}_1$*,* $\mathbb{H}_2$ *be separable Hilbert spaces. Given a measure space* $(E, \mathscr{B}, \mu)$*, for a measurable map* G *on* E *taking values in* $\boldsymbol{\mathcal{B}}_{HS}(\mathbb{H}_1, \mathbb{H}_2)$ *2.2.1 de-*

mands G to be Bochner integrable if,

$$\int_E \|G\|_{HS}\, d\mu < \infty.$$

For any such G,

$$\int_E (Gf) d\mu = \left(\int_E G d\mu \right) f \tag{2.2}$$

for all $f \in \mathbb{H}_1$.

Proof. For any fixed $f \in \mathbb{H}_1$ define a mapping K that maps $G \in \boldsymbol{\mathcal{B}}_{HS}(\mathbb{H}_1, \mathbb{H}_2)$ to $Gf \in \mathbb{H}_2$. We can then rewrite (2.2) as,

$$\int_E K(G) d\mu = K\left(\int_E G d\mu \right). \tag{2.3}$$

Since $K \in \boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{B}}_{HS}(\mathbb{H}_1, \mathbb{H}_2), \mathbb{H}_2)$ and the operator norm of G is bounded above by $\|f\|_1$, meaning (2.3) is simply a result of 2.2.1 □

Theorem 2.2.3. *Suppose that $\mathbb{H}$ is a Hilbert space with inner product and norm $\langle \cdot, \cdot \rangle$, $\|\cdot\|$ and $\mathcal{F} \in \boldsymbol{\mathcal{B}}(\mathbb{H}, \mathbb{R})$. There is a unique element $e_{\mathcal{F}} \in \mathbb{H}$ called the representer of $\mathcal{F}$ with the property,*

$$\mathcal{F}x = \langle x, e_{\mathcal{F}} \rangle,$$

for all $x \in \mathbb{H}$ and $\|\mathcal{F}\| = \|e_{\mathcal{F}}\|$

Definition 2.2.4. *Let $\mathbb{X}$ be an inner-product space with $\mathbb{M} \subset \mathbb{X}$. The orthogonal complement of $\mathbb{M}$ is the set,*

$$\mathbb{M}^{\perp} = \{x \in \mathbb{X} : \langle x, y \rangle = 0,\ \forall x, y \in \mathbb{M}\}.$$

Theorem 2.2.5. *Let $\mathbb{H}$ be a Hilbert space with $\mathbb{M}$ a subset of $\mathbb{H}$. Then,*

(a) $\mathbb{M}^{\perp}$ is a closed subspace.

(b) $\mathbb{M} \subset (\mathbb{M}^{\perp})^{\perp}$.

(c) $(\mathbb{M}^{\perp})^{\perp} = \overline{\mathbb{M}}$ *if* $\mathbb{M}$ *is a subspace.*

Theorem 2.2.6. *Let* $\mathcal{F} \in \boldsymbol{\mathcal{B}}(\mathbb{H}_1, \mathbb{H}_2)$*, for real Hilbert spaces* $\mathbb{H}_1$*,* $\mathbb{H}_2$*. Then,*

(a) $(\mathcal{F}^*)^* = \mathcal{F}$.

(b) $\|\mathcal{F}^*\| = \|\mathcal{F}\|$.

(c) $\|\mathcal{F}^*\mathcal{F}\| = \|\mathcal{F}\|^*$.

(d) $\ker(\mathcal{F}) = (\mathrm{im}(\mathcal{F}^*))^{\perp}$.

(e) $\ker(\mathcal{F}^*\mathcal{F}) = \ker \mathcal{F}$ *and* $\overline{\mathrm{im}(\mathcal{F}^*\mathcal{F})} = \overline{\mathrm{im}(\mathcal{F})^*}$.

Definition 2.2.7. *Let* $x_1 \in \mathbb{H}_1$ *and* $x_2 \in \mathbb{H}_2$*, the tensor product operator* $(x_1 \otimes_1 x_2) : \mathbb{H}_1 \to \mathbb{H}_2$ *is defined by,*

$$(x_1 \otimes_1 x_2)y = \langle x_1, y \rangle_1 x_2,$$

for $y \in \mathbb{H}_1$*. When* $\mathbb{H}_1 = \mathbb{H}_2$ *we simply use* $\otimes$.

Theorem 2.2.8. *Let* $x_1 \in \mathbb{H}_1$ *and* $x_2 \in \mathbb{H}_2$*, then* $\|x_1 \otimes_1 x_2\| = \|x_2\|_2 \|x_1\|_1$.

Proof. For $x_1 \neq 0$ we have,

$$\|x_1 \otimes_1 x_2\| = \sup_{\|v\|=1} \|\langle x_1, v \rangle x_2\|_2 \leqslant \|x_2\|_2 \|x_1\|_1,$$

with equality when $v = x_1 / \|x_1\|_1$. □

Now some common key properties relating to self-adjoint and compact operators.

Theorem 2.2.9. *Let $\mathcal{F}$ be a compact, self-adjoint operator on a Hilbert space $\mathbb{H}$. The set of nonzero eigenvalues for $\mathcal{F}$ is finite or is made up of a sequence tending towards 0. Each nonzero eigenvalue has finite multiplicity and eigenvectors corresponding to different eigenvalues are orthogonal. Let $\lambda_1, \lambda_2, \ldots$ be eigenvalues ordered such that,*

$$|\lambda_1| \geq |\lambda_2| \geq \ldots,$$

and let $e_1, e_2, \ldots$ be the corresponding orthonormal vectors obtained using Gram-Schmidt orthogonalization as necessary for repeated eigenvalues. Then $\{e_i\}$ is an orthonormal basis for $\overline{\operatorname{im} \mathcal{F}}$ and,

$$\mathcal{F} = \sum_{i \geq 1} \lambda_i e_i \otimes e_i.$$

So, for every $x \in \mathbb{H}$,

$$\mathcal{F}x = \sum_{i \geq 1} \lambda_i \langle x, e_i \rangle e_i.$$

Theorem 2.2.10. *Let the continuous kernel K be symmetric and nonnegative definite and $\mathcal{K}$ the corresponding integral operator. If (λ_i, e_i) are eigenvalue and eigenfunction pairs of $\mathcal{K}$, then K has the representation,*

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i e_i(s) e_i(t),$$

for all s, t with the sum converging absolutely and uniformly.

Chapter 3

Functional Data Analysis

3.1 Functional Data

A stochastic process is an indexed collection of random variables which are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let us refer to the indexed set as E. This is simply the following collection,

$$\{X(t,\omega) : t \in E, \omega \in \Omega\},$$

with $X(t,\cdot)$ is a $\mathcal{F}$ measurable function of Ω, and will be shortened to just $X(t)$. Once we observe $X(t)$ for all elements in our index set E, the stochastic process has then been realized, leaving us with a collection of real numbers. This collection is referred to as a sample path for the process. Functional data analysis' (FDA) main concern is the development of methodology and tools for the analysis of data that represent these sample paths, where usually the index set is some closed interval, specifically $[0,1]$. This leads to the analysis of observations that are functions on $[0,1]$ and the data sets are now viewed as a collection of random curves.

We cannot actually observe functional data since we will encounter numerical issues at one point. Due to this fact, analysis usually has to be predicted on n

points through $[0, 1]$ for each sample path, which is finite dimensional data and one could be drawn to do traditional multivariate analysis. This would be fine if the data were not functional data, which is when we have significantly more observations than we do sample paths, multivariate analysis begins to run into many problems as this difference increases [9] which is where the strength of FDA is highlighted.

3.2 Statistics In a Hilbert Space

There are two somewhat different perspectives of functional data. The first is that functional data are realizations of random variables that take values in a Hilbert space; this is the random element perspective. The second view is that functional data is really sample paths of a stochastic process; this is the stochastic process perspective. FDA is the first line of thought, but to begin, one needs to lay a foundation for the study of Hilbert space valued random variables, so we should have concepts such as mean and covariance in this abstract environment.

Let χ be a random element of a separable Hilbert space $\mathbb{H}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. One notion of the mean is the following.

Definition 3.2.1. *If $\mathbb{E}\|\chi\| < \infty$ the mean element of χ, m, or simply the mean of χ is defined as the Bochner integral,*

$$m = \mathbb{E}(\chi) = \int_{\Omega} \chi d\mathbb{P}.$$

This gives us a natural extension of the mean of a random variable to the case of random elements. Roughly speaking, it is a weighted sum of possible realizations of χ that returns another non-random element of $\mathbb{H}$. Naturally, what follows after expectation is a type of variance measure, we define the analog as follows.

Theorem 3.2.2. *Assume that* $\mathbb{E}\|\chi\|^2 < \infty$. *Then,*

$$\mathbb{E}\|\chi - m\|^2 = \mathbb{E}\|\chi\|^2 - \|m\|^2,$$

where m *is the mean of* χ.

Now the next natural step is extending the concept of *covariance*. Recall that for a random $p-$vector X we have its covariance as,

$$\mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T] = \mathbb{E}[(X - \mathbb{E}X) \otimes (X - \mathbb{E}X)].$$

This is a $p \times p$ matrix and therefore an element of $\mathcal{B}(\mathbb{R}^P)$. For Hilbert spaces we build on this idea, if χ is a random element from our Hilbert space, we define the *covariance operator* as follows.

Definition 3.2.3. *Assume that* $\mathbb{E}\|\chi\|^2 < \infty$. *Then the covariance operator for* χ *is the element of* $\mathcal{B}(\mathbb{H})$ *given by the Bochner integral,*

$$\mathcal{K} = \mathbb{E}[(\chi - m) \otimes (\chi - m)] := \int_\Omega (\chi - m) \otimes (\chi - m) d\mathbb{P}.$$

Where m *is the mean of* χ.

Now we provide an extension of a common covariance identity in finite dimensions.

Theorem 3.2.4.

$$\mathbb{E}[(\chi - m) \otimes (\chi - m)] = \mathbb{E}(\chi \otimes \chi) - m \otimes m.$$

Proof. We have $\chi \otimes m$, $m \otimes \chi$, and $m \otimes m$ as HS operators, and $m \otimes m$ specifically is constant in Ω while,

$$\mathbb{E}\|\chi \otimes m\|_{HS} = \mathbb{E}\|m \otimes \chi\|_{HS} = \|m\| \, \mathbb{E}\|\chi\|,$$

so, it holds if for all $g \in \mathbb{H}$,

$$\mathbb{E}(m \otimes \chi)g = \mathbb{E}(\chi \otimes m)g = (m \otimes m)g = \langle m, g \rangle m,$$

which follows from Corollary 2.2.2

□

Generally we assume that $m = 0$ and will state otherwise if not, reducing the above to,

$$\mathcal{K} = \mathbb{E}(\chi \otimes \chi) := \int_{\Omega} (\chi \otimes \chi) d\mathbb{P}.$$

Now we have the following key properties of the covariance operator which will help build towards an analog of PCA,

Theorem 3.2.5. *Assume that $\mathbb{E}\|\chi\|^2 < \infty$. Then for any $f, g \in \mathbb{H}$ we have the following:*

(a) $\langle \mathcal{K}f, g \rangle = \mathbb{E}[\langle \chi, f \rangle \langle \chi, g \rangle]$.

(b) $\mathcal{K}$ *is a nonnegative-definite trace-class operator with,*

$$\|\mathcal{K}\|_{TR} = \mathbb{E}\|\chi\|^2.$$

(c) $\mathbb{P}\left(\chi \in \overline{\operatorname{im} \mathcal{K}}\right) = 1$.

Proof. (a) Notice that $\mathcal{K} \in \boldsymbol{\mathcal{B}}_{HS}(\mathbb{H})$, allowing us to apply Corollary 2.2.2, granting us,

$$\mathcal{K}f = \left(\int_{\Omega} \chi \otimes \chi d\mathbb{P} \right) f = \int_{\Omega} \chi \langle \chi, f \rangle d\mathbb{P},$$

for any $f \in \mathbb{H}$. Then applying Theorem 2.2.1 to the linear functional $Tf := \langle f, g \rangle$ gives us our desired result.

(b) The nonnegative definite property follows from (a), to show that $\mathcal{K}$ is trace class, we let $\{e_i\}_{i\in I}$ be a orthonormal basis for $\mathbb{H}$. Now observe that,

$$\|\mathcal{K}\|_{TR} = \sum_{j=1}^{\infty} \langle \mathcal{K}e_j, e_j \rangle = \sum_{j=1}^{\infty} \mathbb{E}\, \langle \chi, e_j \rangle^2 = \mathbb{E}\, \|\chi\|^2 < \infty.$$

(c) Using (d) of 2.2.6 we have,

$$(\operatorname{im} \mathcal{K})^{\perp} = \ker \mathcal{K}^* = \ker \mathcal{K},$$

since $\mathcal{K}$ is self-adjoint, therefore for any $f \in (\operatorname{im} \mathcal{K})^{\perp}$,

$$\mathbb{E}\left[\langle \mathcal{K}f, f \rangle^2\right] = \langle \mathcal{K}f, f \rangle = 0.$$

Implying that χ is orthogonal to any function in $(\operatorname{im} \mathcal{K})^{\perp}$ with probability one. Then, by part (c) of 2.2.5 we have,

$$\chi \in (\operatorname{im} \mathcal{K})^{\perp\perp} = \overline{\operatorname{im} \mathcal{K}},$$

with probability one.

□

Now with (b) of 3.2.5 and Theorem 2.2.9 we have the foundation of functional PCA,

Theorem 3.2.6. *The covariance operator $\mathcal{K}$ possesses an eigen decomposition,*

$$\mathcal{K} = \sum_{i=1}^{\infty} \lambda_i e_i \otimes e_i.$$

The eigenfunctions $\{e_i\}_i^{\infty}$ create an orthonormal basis for $\overline{\operatorname{im} \mathcal{K}}$, and their respective eigenvalues $\{\lambda_i\}_i^{\infty}$ are non-negative and either finite or a sequence that tends to 0. For nonzero eigenvalues, their multiplicity is finite.

Recalling that a (co)variance matrix of a random vector has a spectral decom-

position, this theorem is an extension of exactly that. Then, like in the finite dimensional case, with this theorem and Theorem 3.2.5 we have an extension of principal component decomposition, which opens the doors to functional PCA.

Theorem 3.2.7. *Assuming that the covariance operator $\mathcal{K}$ has the eigen decomposition in 3.2.6. Then,*

$$\chi = \sum_{i=1}^{\infty} \langle \chi, e_i \rangle \, e_i$$

with probability one, where $\langle \chi, e_i \rangle$ are uncorrelated random variables with mean zero and variances λ_i for all $i \geq 1$.

This is our extension of principal component decomposition stated formally. Similarly to the finite dimensional case, Theorem 3.2.7 provides one with a wide range of uses and properties, one is the following.

Theorem 3.2.8. *Let $\{g_i\}_{i=1}^{\infty}$ be some orthonormal basis for $\mathbb{H}$ then,*

$$\mathbb{E} \left\| \chi - \sum_{i=1}^{n} \langle \chi, g_i \rangle \, g_i \right\|^2 = \mathbb{E} \left\| \chi \right\|^2 - \sum_{i=1}^{n} \langle \mathcal{K} g_i, g_i \rangle$$

which can be minimized by taking $g_i = e_i$ for $i \in \{1, \ldots, n\}$

Proof. We have already,

$$\mathbb{E} \left\| \chi - \sum_{i=1}^{n} \langle \chi, g_i \rangle \, g_i \right\|^2 = \mathbb{E} \left\| \chi \right\|^2 + \underbrace{\mathbb{E} \left\| \sum_{i=1}^{n} \langle \chi, g_i \rangle \, g_i \right\|^2}_{\star} - 2\mathbb{E} \left\langle \chi, \sum_{i=1}^{n} \langle \chi, g_i \rangle \, g_i \right\rangle,$$

where for $\star$ above we have,

$$\mathbb{E} \left\| \sum_{i=1}^{n} \langle \chi, g_i \rangle \, g_i \right\|^2 = \mathbb{E} \left\langle \chi, \sum_{i=1}^{n} \langle \chi, g_i \rangle \, g_i \right\rangle,$$

which leaves,

$$\mathbb{E} \left\| \chi - \sum_{i=1}^{n} \langle \chi, g_i \rangle \, g_i \right\|^2 = \mathbb{E} \left\| \chi \right\|^2 - \mathbb{E} \left\langle \chi, \sum_{i=1}^{n} \langle \chi, g_i \rangle \, g_i \right\rangle,$$

$$= \mathbb{E}\left\|\chi\right\|^2 - \sum_{i=1}^{n} \mathbb{E}\left\langle \chi, g_i \right\rangle^2$$
$$= \mathbb{E}\left\|\chi\right\|^2 - \sum_{i=1}^{n} \left\langle \mathcal{K} g_i, g_i \right\rangle .$$

Then by (b) of Theorem 3.2.5 we have $\|\mathcal{K}\| = \mathbb{E}\left\|\chi\right\|^2$ and then applying Corollary 2.2.2 complete our proof.

□

This establishes the tools needed for principal component decomposition, but that alone does not tell us everything about relationships between variables. One may want to explore dependence between two groups of variables. Recalling though in the finite dimensional case one needs a cross-covariance matrix, and just how we established existence of a covariance operator, we can establish the existence of a cross-covariance operator.

To begin, we first make some amendments to our setting. Suppose we instead have two random elements χ_1, χ_2 defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values from separable Hilbert spaces $\mathbb{H}_1, \mathbb{H}_2$ respectively. Also suppose that $\mathbb{E}\left\|\chi_i\right\|^2 < \infty$ for $i = 1, 2$. As before, assume that their mean is also 0 unless otherwise stated.

Then, we define the cross-covariance operator for χ_1, χ_2 to be defined as the Bochner integral,

$$\mathcal{K}_{12} = \int_{\Omega} (\chi_2 \otimes_2 \chi_1) d\mathbb{P}.$$

The justification for existence follows from the covariance operator. Then for any $\omega \in \Omega$ we have $\chi_2(\omega) \otimes_2 \chi_2(\omega)$ which is a HS operator with norm $\left\|\chi_1(\omega)\right\|_1 \left\|\chi_2(\omega)\right\|_2$, letting us apply Theorem 2.2.1 to show that this integral is well-defined as an element of $\boldsymbol{\mathcal{B}}_{HS}(\mathbb{H}_2, \mathbb{H}_1)$. Now, just as Theorem 3.2.5 demonstrated key properties

of the covariance operator, it is extended to this cross-covariance operator with the following theorem.

Theorem 3.2.9. *Assuming that* $\mathbb{E}\chi_1 = \mathbb{E}\chi_2 = 0$ *and both* $\mathbb{E}\,\|\chi_1\|_1^2$ *and* $\mathbb{E}\,\|\chi_2\|_2^2$ *are finite, then for any* $g \in \mathbb{H}_1$ *and* $f \in \mathbb{H}_2$,

(a) $\langle \mathcal{K}_{12} f, g\rangle_1 = \mathbb{E}\left[\langle \chi_1, g\rangle_1 \langle \chi_2, f\rangle_2\right]$.

(b) $|\langle \mathcal{K}_{12} f, g\rangle_1| \leqslant \langle \mathcal{K}_1 g, g\rangle_1^{1/2} \langle \mathcal{K}_2 f, f\rangle_2^{1/2}$.

(c) $\mathcal{K}_{12}$ *has adjoint,*

$$\mathcal{K}_{21} = \int_\Omega (\chi_1 \otimes_1 \chi_2) d\mathbb{P}.$$

Proof. Parts (b) and (c) follow from (a), and we have (a) in the same fashion we have (a) of Theorem 3.2.5. □

In traditional multivariate analysis, the generalized correlation measure is provided by the matrix,

$$\mathcal{R}_{12} = \mathcal{K}_1^{1/2} \mathcal{K}_{12} \mathcal{K}_2^{1/2},$$

where $\mathcal{K}_1$, $\mathcal{K}_2$, and $\mathcal{K}_{12}$ are the covariance and cross-covariance matrices for the two random variables of interest. Canonical correlation usually centers around the singular value decomposition of this measure. Following this pattern of extension, we would like there to be an extension of this in our Hilbert space of random elements, but unfortunately, the compactness of our operators makes them non-invertible in infinite dimensions.

Still though, we can at least establish an analog of the measure $\mathcal{R}$ due to Baker [5].

Theorem 3.2.10. *There exists an operator* $\mathcal{R}_{12} \in \boldsymbol{\mathcal{B}}(\mathbb{H}_1, \mathbb{H}_2)$ *with* $\|\mathcal{R}_{12}\| \leqslant 1$ *such that* $\mathcal{K}_{12} = \mathcal{K}_1^{1/2} \mathcal{R}_{12} \mathcal{K}_2^{1/2}$.

Proof. First, let (λ_{1i}, e_{1i}) be eigenvalues and eigenfunctions of $\mathcal{K}_1$ and $\mathcal{P}_n$ the projection in $\mathbb{H}_1$ of $\text{span}\{e_{1i}, \ldots, e_{1n}\}$. Then, we have for every $f \in \mathbb{H}_2$,

$$\begin{aligned}
\left\|\mathcal{P}_n\mathcal{K}_1^{-1/2}\mathcal{K}_{12}f\right\|_1^2 &= \left\langle \mathcal{K}_{12}f, \mathcal{P}_n\mathcal{K}_1^{-1}\mathcal{K}_{12}f\right\rangle_1 \\
&\leqslant \langle \mathcal{K}_2 f, f\rangle_2^{1/2} \left\langle \mathcal{K}_1\mathcal{P}_n\mathcal{K}_1^{-1}\mathcal{K}_{12}f, \mathcal{P}_n\mathcal{K}_1^{-1}\mathcal{K}_{12}f\right\rangle_1^{1/2} \quad \text{by (b) of 3.2.9} \\
&= \langle \mathcal{K}_2 f, f\rangle_2^{1/2} \left\|\mathcal{P}_n\mathcal{K}_1^{-1/2}\mathcal{K}_{12}f\right\|_1 \\
\left\|\mathcal{P}_n\mathcal{K}_1^{-1/2}\mathcal{K}_{12}f\right\|_1 &\leqslant \langle \mathcal{K}_2 f, f\rangle_2^{1/2} \qquad \text{by dividing} \\
\left\|\mathcal{P}_n\mathcal{K}_1^{-1/2}\mathcal{K}_{12}f\right\|_1 &\leqslant \left\|\mathcal{K}_2^{1/2}f\right\|_2,
\end{aligned}$$

which we have for any n. Thus, for every $f \in \mathbb{H}_2$ we have,

$$\left\|\mathcal{K}_1^{-1/2}\mathcal{K}_{12}f\right\|_1 \leqslant \left\|\mathcal{K}_2^{1/2}f\right\|_2.$$

Then, if $f \in \text{im}\left(\mathcal{K}_2^{1/2}\right)$ in that $f = \mathcal{K}_\in{}^{1/2} f_0$ for some $f_0 \in \mathbb{H}_2$ we obtain,

$$\|\underbrace{\mathcal{K}_1^{-1/2}\mathcal{K}_{12}\in^{-1/2}f}_{\mathcal{R}_{12}}\|_1 \leqslant \|f\|_2,$$

meaning that $\mathcal{R}_{12}$ is bounded on $\text{im}\left(\mathcal{K}_2^{1/2}\right)$ with norm at most 1. Now, we can extend $\mathcal{R}_{12}$ to the closure through the extension principle [3], with the extension having the same norm. Lastly, we define $\mathcal{R}_{12}f = 0$ for f in the closure of $Im(\mathcal{K}_2^{1/2})^\perp$ to complete the definition of $\mathcal{R}_{12}$ on all of $\mathbb{H}_1$, as desired.

□

As in the finite dimensional case, this operator gives us a measure of dependence between our random variables χ_1 and χ_2.

When $\|\chi_1, g\|_1$ and $\|\chi_2, f\|_2$ are bivariate normal for all $g \in \mathbb{H}_1$ and $f \in \mathbb{H}_2$, the mutual information (how much information one variable says about the other) of χ_1 and χ_2 is finite if and only if $\mathcal{R}_{12}$ is HS and its norm is strictly less than 1

[5]. This concludes our extensions of common statistical tools and objects to the Hilbert space setting.

3.3 Stochastic Process

We now explore the other line of thought which is viewing functional data as a stochastic process. Our setting is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with stochastic process $X = \{X(t) : t \in E\}$. Where X is representing a random function that is partially observed or realized. Under measure-theoretic assumptions of a stochastic process, $X(t)$ is a random variable, but this does not demand $X(\cdot)$ to be a random element of $L^2(E, \mathscr{B}(E), \mu)$. Because of this, conditions have been established as to when this is guaranteed [2]. With this established, we begin our exploration. The mean function of the stochastic process X is defined as,

$$m(t) = \mathbb{E}[X(t)],$$

and the covariance kernel as,

$$K(s,t) = Cov(X(s), X(t)),$$

for $s, t \in E$.

Definition 3.3.1. *A second order process is defined as a stochastic process, X, with well-defined mean function and covariance kernel.*

It is clear that the covariance kernel K is nonnegative definite simply by definition of covariance.

Definition 3.3.2. *A mean square continuous process is defined as a stochastic*

process such that,

$$\lim_{n\to\infty} \mathbb{E}[X(t_n) - X(t)]^2 = 0.$$

Theorem 3.3.3. *Let X be a second order process. Then, X is mean square continuous if and only if its mean function and covariance kernel are continuous.*

Proof. ($\Rightarrow$) First we show continuity of the mean function

$$|m(s) - m(t)| = |\mathbb{E}[X(s) - X(t)]| \leqslant \left(\mathbb{E}[X(s) - X(t)]^2\right)^{1/2} \to 0,$$

which follows from our assumption of Definition 3.3.2. Now we assume that $m(t) \equiv 0$, covariance kernel continuity follows from,

$$K(s,t) - K(s',t') = (K(s,t) - K(s',t)) + (K(s',t) - K(s',t'))$$

applying Cauchy-Shwarz inequality we have,

$$|K(s,t) - K(s',t)| \leqslant K^{1/2}(t,t)(\mathbb{E}\left[X(s) - X(s')\right]^2)^{1/2},$$
$$|K(s',t) - K(s',t')| \leqslant K^{1/2}(s',s')(\mathbb{E}\left[X(t) - X(t')\right]^2)^{1/2},$$

which, as with the mean function, follow from our assumption of Definition 3.3.2.

($\Leftarrow$) By definition we have,

$$\mathbb{E}\left[X(s) - X(t)\right]^2 = K(s,s) + K(t,t) - 2K(s,t) + (m(s) - m(t))^2,$$

and the continuity of the mean function and covariance kernel imply Definition 3.3.2. □

Just because X is a mean squared continuous process does not mean it is always a random element of any Hilbert space. Regardless, we have the following integral

operator on $L^2(E, \mathscr{B}(E), \mu)$ which is well-defined,

$$(\mathcal{K}f)(t) = \int_E K(t,s)f(s)d\mu(s),$$

where μ is of finite measure. This $\mathcal{K}$ is the covariance operator of X. Then by Theorem 2.2.10,

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i e_i(s)e_i(t).$$

Now, in the event that X is also a random element of $\mathbb{H}$ we have the following.

Theorem 3.3.4. *Let $X = \{X(t) : t \in E\}$ be a mean square continuous process that is jointly measurable. Then,*

(a) The mean function m belongs to $\mathbb{H}$ and coincides with the mean element of X in $\mathbb{H}$.

(b) The covariance operator $\mathbb{E}(X \otimes X)$ is defined and coincides with the operator $\mathcal{K}$.

(c) For any $f \in \mathbb{H}$,

$$\mathbb{I}_X(f) = \int_E X(t)f(t)d\mu(t) = \langle X, f \rangle .$$

Proof. Refer to [4]. □

We end the examination of stochastic processes here, since what we have so far is sufficient for our purposes.

3.4 Sampling

Now, we explore large-sample results that will be of use in inference problems under FDA. Assume that one has independent realizations $X_1, X_2, \ldots, X_n$ of some

real valued variable X with expectation m. In many situations we have that,

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

converges almost surely to m (Strong law of large numbers), and after normalization, it roughly has a Gaussian distribution (Central limit theorem). It would be wonderful to have extensions of these essential results for when X_i is a random element of a Hilbert space, fortunately we do!

Now we let $\chi_1, \chi_2, \ldots$ be random elements in $\mathbb{H}$, and we denote the sum of the first n as,

$$S_n = \sum_{i=1}^{n} \chi_i$$

Theorem 3.4.1. *Assuming that the collection $\{\chi_i\}_{i=1}^n$ is pairwise independent with $m = 0$, we have,*

$$\mathbb{E}\left\|S_n\right\|^2 = \sum_{i=1}^{n} \mathbb{E}\left\|\chi_i\right\|^2.$$

Proof. Take $\{e_k : k \geq 1\}$ to be an orthonormal basis for $\mathbb{H}$, since we have pairwise independence in the collection $\{\chi_i\}$,

$$\mathbb{E}\left\langle \chi_i, e_k\right\rangle \left\langle \chi_j, e_k\right\rangle = \mathbb{E}\left\langle \chi_i, e_k\right\rangle \mathbb{E}\left\langle \chi_j, e_k\right\rangle = \left\langle \mathbb{E}\chi_i, e_k\right\rangle \left\langle \mathbb{E}\chi_j, e_k\right\rangle = 0,$$

when $i \neq j$. Proceeding we have,

$$\begin{aligned}\mathbb{E}\left\|S_n\right\|^2 = \sum_{k=1}^{\infty} \mathbb{E}\left\langle S_n, e_k\right\rangle^2 = \sum_{k=1}^{\infty}\sum_{i=1}^{n} \mathbb{E}\left\langle \chi_i, e_k\right\rangle^2 &= \sum_{i=1}^{n}\sum_{k=1}^{\infty} \mathbb{E}\left\langle \chi_i, e_k\right\rangle^2 \\ &= \sum_{i=1}^{n} \mathbb{E}\left\|\chi_i\right\|^2.\end{aligned}$$

□

Now we provide an extension of the Strong law of large numbers.

Theorem 3.4.2. *Assume $\chi_1, \chi_2, \ldots$ be pairwise independent, and independent and identically distributed (i.i.d.) with $\mathbb{E}\|\chi_1\| < \infty$. Then,*

$$\lim_{n\to\infty} \frac{1}{n} S_n = \mathbb{E}\chi_1,$$

almost surely.

The outline of this proof follows from Etemadi's clever proof of Strong law of large numbers [4].

Proof. Define,

$$\chi_i' = \chi_i \mathbb{I}_{(\|\chi_i\| \leqslant i)},$$

where $\mathbb{I}$ is the indicator function. Similar to before we also define,

$$S_n' = \sum_{i=1}^{n} \chi_i'.$$

Let $[\alpha]$ denote the integer part of α. Now, take $k_n = [\alpha^n]$ for $\alpha > 1$ and $\{e_k\}_{k=1}^{\infty}$ to be an orthonormal basis for $\mathbb{H}$. We now apply Theorems 3.2.2 and 3.4.1, and Markov's [1] inequality to obtain,

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}\left(\varepsilon < \frac{\left\|S_{k_n}' - \mathbb{E}S_{k_n}'\right\|}{k_n}\right) &\leqslant \varepsilon^{-2} \sum_{n=1}^{\infty} k_n^{-2} \sum_{i=1}^{k_n} \mathbb{E}\left\|\chi_i'\right\|^2 \\ &= \varepsilon^{-2} \sum_{i=1}^{\infty} \mathbb{E}\left\|\chi_i'\right\|^2 \sum_{n:k_n \geq i} k_n^{-2}.\end{aligned}$$

Then following the results from [11] chapter 2 we get a last inequality,

$$\varepsilon^{-2} \sum_{i=1}^{\infty} \mathbb{E}\left\|\chi_i'\right\|^2 \sum_{n:k_n \geq i} k_n^{-2} \leqslant 4(1-\alpha^{-2})^{-1} \varepsilon^{-2} \sum_{i=1}^{\infty} \mathbb{E}\left\|\chi_i'\right\|^2 i^{-2}.$$

We have though,

$$\sum_{i=1}^{\infty} \mathbb{E}\left\|\chi_i'\right\|^2 i^{-2} = \sum_{i=1}^{\infty} i^{-2} \sum_{j=0}^{i-1} \mathbb{E}\left[\left\|\chi_i\right\|^2 \mathbb{I}_{(j \leqslant \|\chi_1\| \leqslant j+1)}\right] \tag{3.1}$$

$$= \sum_{j=0}^{\infty} \mathbb{E}\left[\|\chi_1\|^2 \mathbb{I}_{(j\leqslant\|\chi_1\|\leqslant j+1)}\right] \sum_{i=j+1}^{\infty} i^{-2} \tag{3.2}$$

$$\leqslant C \sum_{j=0}^{\infty} (j+1)^{-1} \mathbb{E}\left[\|\chi_1\|^2 \mathbb{I}_{(j\leqslant\|\chi_1\|\leqslant j+1)}\right], \tag{3.3}$$

for a $C < \infty$. We have (3.3) bounded by $\mathbb{E}[\chi_1]$, so we can apply Borel-Cantelli [1] and see that,

$$\lim_{n\to\infty} \frac{S'_{k_n} - \mathbb{E}S'_{k_n}}{k_n} = 0,$$

almost surely. Then, applying Lebesgue's dominated convergence theorem,

$$\begin{aligned}
&\lim_{n\to\infty} \|\mathbb{E}\chi'_n - \mathbb{E}\chi_1\| \leqslant \lim_{n\to\infty} \mathbb{E}\,\|\chi_1\|\,\mathbb{I}_{(n<\|\chi_1\|)} = 0 \\
\implies &\lim_{n\to\infty} \left\| \frac{\mathbb{E}S'_{k_n}}{k_n} - \mathbb{E}\chi_1 \right\| \leqslant \lim_{n\to\infty} \frac{1}{k_n} \sum_{i=1}^{k_n} \|\mathbb{E}\chi'_i - \mathbb{E}\chi_1\| = 0 \\
\implies &\lim_{n\to\infty} \frac{S'_{k_n}}{k_n} = \mathbb{E}\chi_1,
\end{aligned}$$

almost surely. Then, with Borel-Cantelli, we have that $X'_n = X_n$ eventually with probability 1,

$$\lim_{n\to\infty} \frac{S_{k_n}}{k_n} = \mathbb{E}\chi_1, \tag{3.4}$$

almost surely. Now, we have $m(n) \in \mathbb{N}$ such that,

$$k_{m(n)-1} = [\alpha^{m(n)-1}] < n \leqslant [\alpha^{m(n)}] = k_{m(n)}.$$

Using this, we have

$$\begin{aligned}
\left\| \frac{S_n}{n} - \frac{S_{k_{m(n)}}}{k_{m(n)}} \right\| &= \left\| \frac{S_{k_{m(n)}}}{n} - \frac{S_{k_{m(n)}}}{k_{m(n)}} + \frac{S_n - S_{k_{m(n)}}}{n} \right\| \\
&\leqslant \left(\frac{k_{n(n)}}{n} - 1 \right) \left\| \frac{S_{k_{m(n)}}}{k_{m(n)}} \right\| + \frac{1}{n} \sum_{i=n+1}^{k_{m(n)}} \|\chi_i\|
\end{aligned}$$

$$\leqslant (\alpha - 1) \left\| \frac{S_{k_{m(n)}}}{k_{m(n)}} \right\| + \frac{1}{n} \sum_{i=n+1}^{k_{m(n)}} \|\chi_i\| .$$

Now, with (3.4) we have,

$$\lim_{n\to\infty} \left\| \frac{S_{k_{m(n)}}}{k_{m(n)}} \right\| = \|\mathbb{E}\chi_1\| \leqslant \mathbb{E}\,\|\chi_1\| ,$$

almost surely. Then, by the standard Strong law of large numbers for real values we are guaranteed that both,

$$\underbrace{\frac{1}{k_{m(n)}} \sum_{i=1}^{k_{m(n)}} \|\chi_i\|}_{\heartsuit}, \underbrace{\frac{1}{n} \sum_{i=1}^{n} \|\chi_i\|}_{\clubsuit} \to \mathbb{E}\,\|\chi_1\| ,$$

with probability 1. Meaning,

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \sum_{i=n+1}^{k_{m(n)}} \|\chi_i\| &= \limsup_{n\to\infty} \left(\frac{k_{m(n)}}{n} \heartsuit - \clubsuit \right) \\ &\leqslant (\alpha - 1)\mathbb{E}\,\|\chi_1\| , \end{aligned}$$

almost surely, and thereby giving us,

$$\limsup_{n\to\infty} \left\| \frac{S_n}{n} - \frac{S_{k_{m(n)}}}{k_{m(n)}} \right\| \leqslant 2(\alpha - 1)\mathbb{E}\,\|\chi_1\| ,$$

and as one lets $\alpha \downarrow 1$ we get our desired result. □

In all, we now have an analog of the Strong law of large numbers in a Hilbert space setting. Now all that is left is to get an analog of the Central limit theorem, which is just convergence in distribution. To do so, we need a notion of weak convergence for probability measures.

Definition 3.4.3. *Let $\mathbb{P}, \mathbb{P}_n$ for $n \geq 1$ be probability measures on $(\mathbb{H}, \mathscr{B}(\mathbb{H}))$. $\mathbb{P}_n$*

converges weakly to $\mathbb{P}$ *if,*

$$\int_{\mathbb{H}} f(x)d\mathbb{P}_n(x) \to \int_{\mathbb{H}} f(x)d\mathbb{P}(x),$$

for any bounded and continuous functions f *on* $\mathbb{H}$*, this is denoted as* $\mathbb{P}_n \xrightarrow{p} \mathbb{P}$

For random elements χ, χ_n for $n \geq 1$ we say that χ_n converges in distribution if,

$$\mathbb{P} \circ \chi_n^{-1} \to \mathbb{P} \circ \chi^{-1},$$

and denote it by $\chi_n \xrightarrow{d} \chi$ [1].

Now we lay out some necessary tools in order to get our Central limit theorem.

Definition 3.4.4. *An arbitrary set of probability measures* $\{\mu_\alpha\}_{\alpha \in I}$ *on* $(\mathbb{H}, \mathscr{B}(\mathbb{H}))$ *is tight if for any* $\varepsilon > 0$ *there exists a compact set* W *such that,*

$$\inf_{\alpha \in I} \mu_\alpha(W) \geq 1 - \varepsilon.$$

For any $S \subset \mathbb{H}$ *and any* $\varepsilon > 0$*, let,*

$$S^c = \{x \in \mathbb{H} \mid \inf\{\|x - z\| : z \in S\} \leqslant \varepsilon\}.$$

Theorem 3.4.5. *Let* $\{\mu_\alpha\}_{\alpha \in I}$ *be a family of probability measures on* $(\mathbb{H}, \mathscr{B}(\mathbb{H}))$*. Assume that for each* $\varepsilon, \delta > 0$ *there exists a finite subset* $\{y_1, \ldots, y_k\} \subset \mathbb{H}$ *such that,*

(a) $\inf_{\alpha \in I} \mu_\alpha(S^c) \geq 1 - \delta$ *where* $S := span\{y_1, \ldots, y_k\}$*.*

(b) $\inf_{\alpha \in I} \mu_\alpha(\{x \in \mathbb{H} : |\langle x, y_j \rangle| \leqslant r, j = 1, \ldots, k\}) \geq 1 - \delta$ *for some* $r > 0$*.*

Then, $\{\mu_\alpha\}_{\alpha_I}$ *is tight.*

Proof. Refer to [1] chapter 4. □

Theorem 3.4.6. *Let χ, χ_n for $n \geq 1$ be random elements in $(\mathbb{H}, \mathscr{B}(\mathbb{H}))$. Assume that $\langle \chi_n, f \rangle \to \langle \chi, f \rangle$ in $\mathbb{R}$ for all $f \in \mathbb{H}$ and for each $\varepsilon, \delta > 0$, there exists a finite dimensional subspace S such that,*

$$\inf_{n \geq 1} \mathbb{P}(\chi_n \in S^c) \geq 1 - \delta,$$

for S^c defined as in Theorem 3.4.5. *Then $\chi_n \xrightarrow{d} \chi$.*

Proof. Refer to [1] chapter 4. □

Now, we construct a Central limit theorem analog for random elements in a Hilbert space.

Theorem 3.4.7. *Let $\chi_1, \chi_2, \ldots$ be i.i.d. random elements in $\mathbb{H}$ with mean 0 and $\mathbb{E}\,\|\chi_1\|^2 < \infty$. Then,*

$$\zeta_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \chi_i \xrightarrow{d} \zeta,$$

where ζ is a Gaussian random element of $\mathbb{H}$ with covariance operator equal to $\mathbb{E}(\chi_1 \otimes \chi_1)$.

Proof. This will follow from Theorem 3.4.6. First, by the standard Central limit theorem for reals, for any $f \in \mathbb{H}$ the distribution of $\langle \zeta_n, f \rangle$ converges to $N(0, \langle \mathbb{E}(\chi_1 \otimes \chi_1), f \rangle)$, which is the distribution of $\langle \zeta, f \rangle$.

Then, let $\{e_j\}$ be an orthonormal basis for $\mathbb{H}$ and let $S_K = span\,\{e_1, \ldots, e_K\}$ be the S from Theorem 3.4.5. Let ζ_{nK} and ζ'_{nK} be the projections of ζ_n on S_K and $S_K^{\perp}$ respectively, and let χ_{iK} be the projection of χ_i on $S_K^{\perp}$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}(\|\zeta'_{nK}\| \leqslant \varepsilon) = \mathbb{P}(\zeta_n \in S_K^c).$$

Then by Chebyshev's inequality,

$$\mathbb{P}(\|\zeta'_{nK} > \varepsilon\|) \leqslant \varepsilon^{-1} \mathbb{E}(\|\chi'_{1K}\|^2),$$

which will be smaller than any δ for K sufficiently large. This satisfies both conditions for Theorem 3.4.6, giving us our Central limit theorem. $\square$

Now, we will discuss actually estimating these objects. As before, let $\mathbb{H}$ be a separable Hilbert space and χ be a random element of said space. Assume that one observed i.i.d. samples $\chi_1, \ldots, \chi_n$ from χ. We want to estimate the mean, m, and the covariance operator, $\mathcal{K}$. Similar to traditional methods, we define our sample mean and covariance operator as follows

$$m_n = \frac{1}{n}\sum_{i=1}^{n} \chi_i$$

and

$$\mathcal{K}_n = \frac{1}{n-1}\sum_{i=1}^{n} (\chi_i - m_n) \otimes (\chi_i - m_n).$$

Now applying our sampling results, we obtain the following immediately about our sample mean.

Theorem 3.4.8. *If* $\mathbb{E}\,\|\chi_1\| < \infty$ *then* $m_n \to m$ *almost surely. If* $\mathbb{E}\,\|\chi_1\|^2 < \infty$ *then we have in* $\mathbb{H}$*,*

$$\sqrt{n}(m_n - m) \xrightarrow{d} \zeta,$$

where ζ *is Gaussian with mean* 0*.*

We also have the following asymptotic properties of the sample covariance operator.

Theorem 3.4.9. *If* $\mathbb{E}\,\|\chi_1\|^2 < \infty$ *then* $\mathcal{K}_n \to \mathcal{K}$ *almost surely. If* $\mathbb{E}\,\|\chi_1\|^4 < \infty$ *then in* $\boldsymbol{\mathcal{B}}_{HS}(\mathbb{H})$*,*

$$\sqrt{n}(\mathcal{K}_n - \mathcal{K}) \xrightarrow{d} \beta,$$

where β is our Gaussian random element with mean zero and covariance operator,

$$\mathbb{E}((\chi_1 - m) \otimes (\chi_1 - m) - \mathcal{K}) \otimes_{HS} ((\chi_1 - m) \otimes (\chi_1 - m) - \mathcal{K}).$$

Proof. First,

$$\mathcal{K}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\chi_i - m) \otimes (\chi_i - m) - \frac{n}{n-1} (m_n - m) \otimes (m_n - m),$$

where the former conclusion follows directly from our Strong law of large numbers 3.4.2 and Theorem 3.4.9. The latter follows from our Central limit theorem 3.4.7 if,

$$\mathbb{E} \left\| (\chi_i - m) \otimes (\chi_i - m) - \mathcal{K} \right\|_{HS}^2 < \infty.$$

With Theorem 3.2.2 we have the following though,

$$\mathbb{E} \left\| (\chi_i - m) \otimes (\chi_i - m) - \mathcal{K} \right\|_{HS}^2 \leqslant \mathbb{E} \left\| (\chi_i - m) \otimes (\chi_i - m) \right\|_{HS}^2 = \mathbb{E} \left\| \chi_i - m \right\|^4,$$

and the right most equality is clearly finite, giving us our desired result. □

This concludes exploration of the purer side of FDA. Although everything explored to this point cannot be fully observed in practice due to the nature of infinity, the results will still be used with what follows in the last section.

3.5 Confidence

In traditional statistics and in finite dimensions, once sampling is done, a sample mean and sample covariance matrix is calculated. From there, usually there is a desire to make an inference on the whole population from the sample, specifically how the sample mean generalizes to the population/true mean. This is done through the Central limit theorem to create a confidence interval. One can try to

replicate this for sample paths,

$$X_i(t_j) = Y_i(t_j) + \varepsilon_i(t_j),$$

where i is the ith sample path, and t_j is our time index. In this situation we model X_i through what is believed to be its true underlying function Y_i with some noise ε_i from the observation itself. Calculating the sample mean as,

$$m(t_j) = \frac{1}{n}\sum_{i=1}^{n} X_i(t_j).$$

This is still discrete, as mentioned at the start, so some form of interpolation is performed to turn these discrete sample paths into actual curves (Fourier, B-splines, Monomial). So, we will have a curve for each Y_i based on whatever choice of interpolation, and then a curve for the mean function m based on the same interpolation as well.

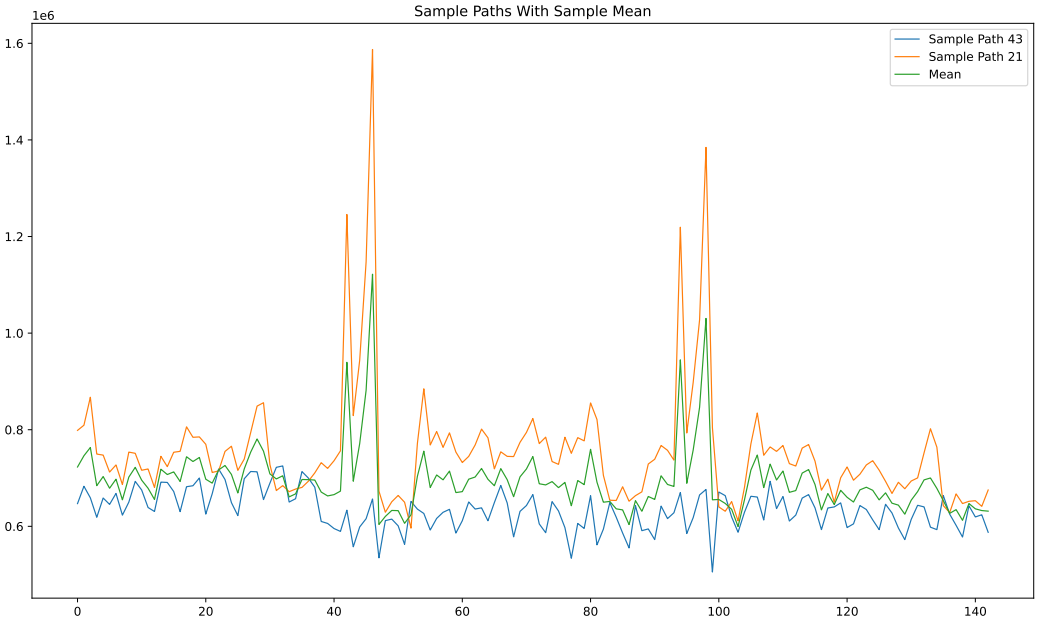

Figure 3.1: Two sample paths and mean.

If we took the confidence interval of the mean function (before interpolation)

at each time sample, of course the mean will always be between the confidence interval at each time step as seen in Figure 3.2. One might then have the desire

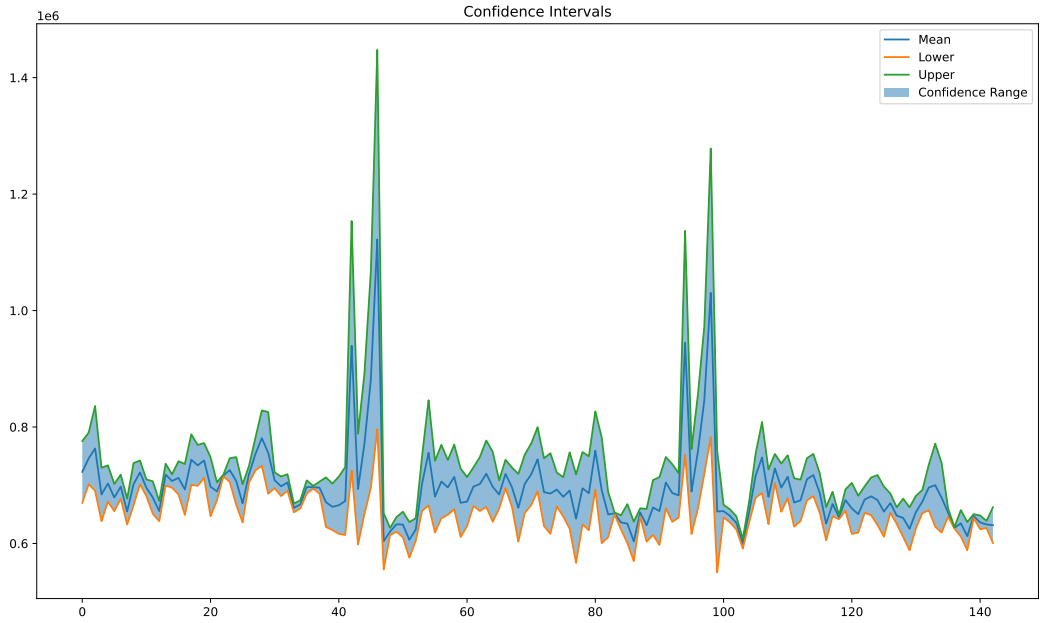

Figure 3.2: Confidence band for each time step.

to interpolate the mean and the confidence intervals under the same interpolation to generate a type of confidence band. One will stumble onto the issue though that the continuous mean function in between time steps can sometimes travel outside these confidence bands as seen a few times in Figure 3.3, in that case even the upper limit confidence band goes below the lower limit confidence band which makes no real world sense. The reason this option is explored in the first place is that even though we have a functional Central limit theorem, generating confidence bands for our sample mean function that makes sense cannot be done analytically. Many refer to proper confidence bands as simultaneous confidence bands, since a confidence interval must make sense for any point in time on the continuous space, not just at the discrete time steps. There is much effort put into coming up with ways to obtain these simultaneous confidence bands using other

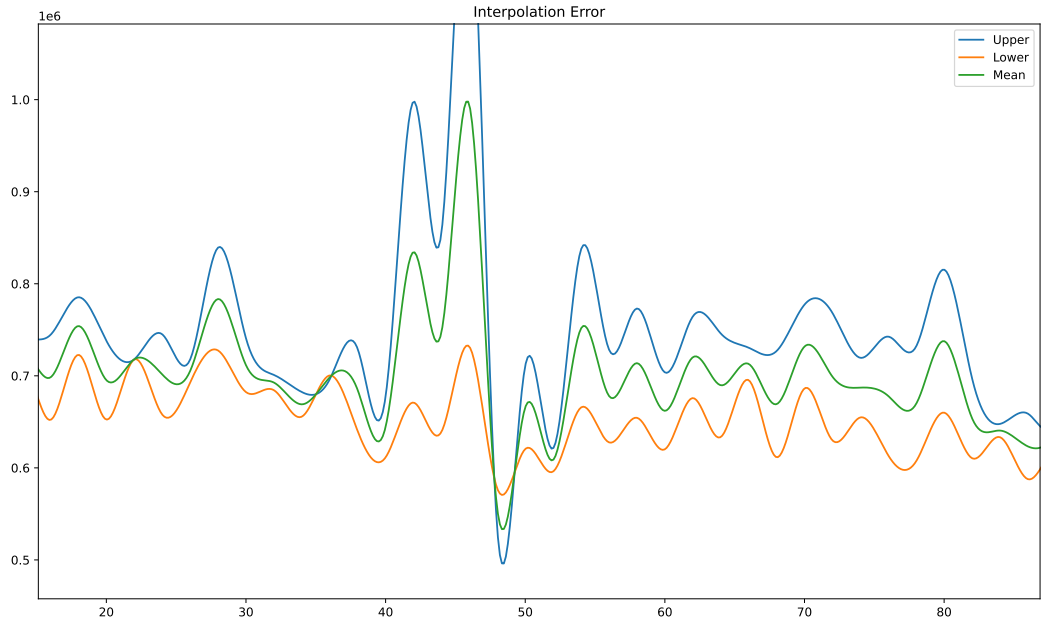

Figure 3.3: Mean function traveling outside of confidence bands.

statistical methods such as Bootstrapping [7] or implementing Bayesian inference [8]. We will not dive into these methods here, but it is worth mentioning.

3.6 Functional PCA

Recall in the finite dimensional case, when one has a p-dimensional random vector X with expectation m and covariance matrix $\mathcal{K}$, if,

$$\mathcal{K} = \sum_{i=1}^{p} \lambda_i e_i e_i^T,$$

is the eigen-value/vector decomposition for the covariance matrix, one could decompose X as,

$$X = m + \sum_{i=1}^{p} Z_i e_i.$$

Where the random variables $Z_i = e_i^T(X - m)$ are of mean zero and uncorrelated with variance λ_i. These random variables are the principal components of X.

Each of these principal components explain a portion of variance from the original data, meaning if one is willing to sacrifice some portion of variance and only use n components (where $n < p$) then one can form a substitute of X (within their tolerance) and achieve dimensional reduction in the process.

We can achieve this dimensional reduction in the infinite dimensional setting as well, through what we have built up, and specifically the functional analog of principal components.

Now translating our workspace, suppose that χ is a random element of a Hilbert space $\mathbb{H}$ instead and $\mathbb{E}\left\|\chi\right\|^2 < \infty$. Now with Theorem 3.2.6, we know that χ has a covariance operator $\mathcal{K}$ that admits the spectral decomposition,

$$\mathcal{K} = \sum_{i=1}^{\infty} \lambda_i e_i \otimes e_i.$$

We also know from Theorem 3.2.7 that,

$$\chi = \mathbb{E}\chi + \sum_{i=1}^{\infty} \langle (\chi - \mathbb{E}\chi), e_i \rangle = m + \sum_{i=1}^{\infty} Z_i e_i,$$

where Z_i are of mean zero and uncorrelated random variables, with variance λ_i. One might be able to sniff out now that the multivariate analysis equivalent of PCA is merely a case of when our Hilbert space is $\mathbb{R}^p$!

Now, when we have a stochastic process $X = \{X(t) : t \in E\}$ that also happens to be a random element of a Hilbert space $\mathbb{H}$, functional PCA emerges. By theorem 3.3.4, knowledge of the covariance operator $\mathcal{K}$ is equivalent to knowing the process covariance kernel,

$$K(s,t) = Cov(X(s), X(t)),$$

and by Theorem 2.2.10, when X is mean-square continuous,

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i e_i(s) e_i(t),$$

with $\{(\lambda_i, e_i)\}_{i=1}^{\infty}$ being the eigen sequence for the $\mathbb{H}$ integral operator corresponding to K, and that the sequence converges absolutely and uniformly in s and t. We have by Karhunen-Loeve expansion [7],

$$X(t) = m(t) + \sum_{i=1}^{\infty} Z_i e_i(t),$$

with $m(t) = \mathbb{E}X(t)$ and,

$$Z_i = \int_E (X(t) - m(t)) e_i(t) d\mu(t).$$

Which is our functional principal component decomposition analog. By truncating the infinite sum to the first N terms, for a large enough N, one can obtain a good approximation of the infinite sum, and thereby creating a substitute of X that captures most of its information, but with only N components. In the infinite dimensional case, the effect of dimensional reduction is significantly more powerful, which is why much effort is put into showing existence and finding principal components.

Chapter 4

Applications

4.1 Introduction

First, recall that functional data is generally considered data that has significantly more observations than sample paths. So, the dataset for our application will be on the weekly sales for 50 stores over 3 years. In this case we have 156 observations for 50 sample paths as seen in Figure 4.1. We will be doing this application in Python, using Pandas, NumPy, and scikit-learn libraries. Our aim will be to forecast weekly sales based on given data. To do so, we will conduct regression through functional PCA.

4.2 Forecasting

First, we turn our discrete sample paths into continuous curves that live in $L^2[0, 1]$. Our method of choice will be both BSpline (Figure 4.2) and Fourier interpolation to compare performance. Of these store curves though, we will only use 80 percent of them as our sample since this will be our training data, and we can test on the unused 20 percent. We generate a sample covariance operator and obtain the top

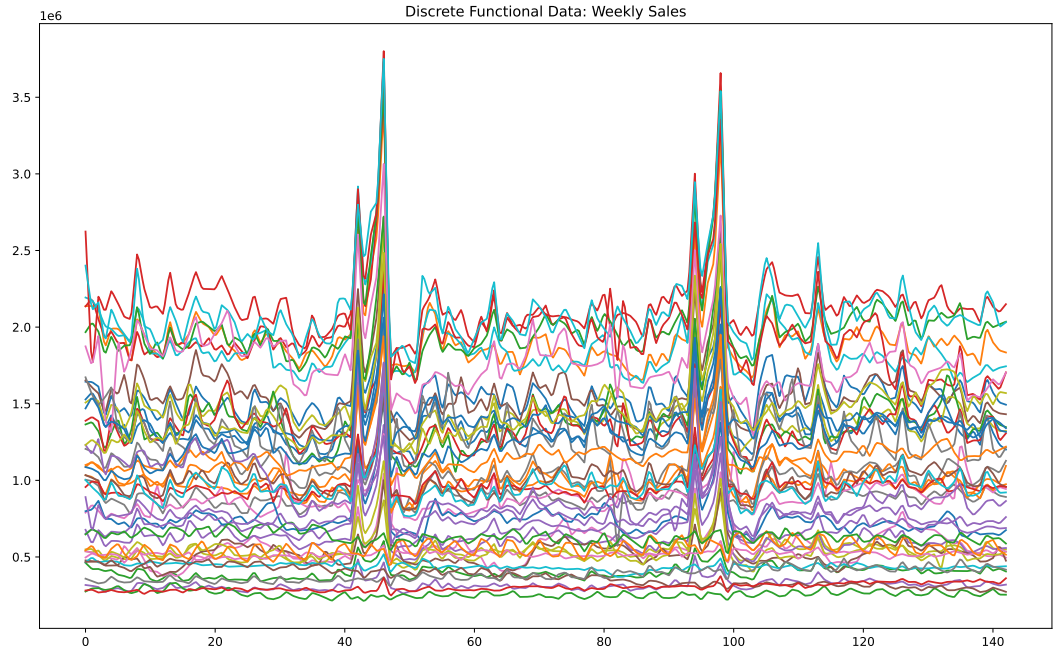

Figure 4.1: Sample paths weekly sales for 50 stores.

three principal components as seen in Figure 4.3.

Where fPC 1, 2, and 3 of Figure 4.3 have an explained variance of $4.52 \cdot 10^{13}$, $2.83 \cdot 10^{11}$, and $1.52 \cdot 10^{11}$, with the explained variance ratios being 0.983, 0.0062, and 0.0033 respectively. With fPC 1 being the best followed by 2 and 3. This means these three curves capture about 99 percent of the variance in the data and should be able to model the data well. Let Z_1, Z_2, and Z_3 be the functional components respectively. Now, we want to create a function using all these functional components,

$$Y(t) = \beta_1 Z_1(t) + \beta_2 Z_2(t) + \beta_3 Z_3(t) + \beta_4,$$

where β_i are optimal constants such that Y fits a given sample of data best under the least squared error measurement. If one is using more principal components, it would be best to use an optimization algorithm such as Gradient Descent, but since we only need to optimize four constants, we can do this analytically. Let

Figure 4.2: BSpline interpolation of discrete data.

$A(t)$ be the actual sales for time t. We want to minimize the following,

$$E = \frac{1}{N}\sum_{i=1}^{N}(\beta_1 Z_1(t_i) + \beta_2 Z_2(t_i) + \beta_3 Z_3(t_i) + \beta_4 - A(t_i))^2,$$
$$E = \frac{1}{N}\sum_{i=1}^{N}(Y(t_i) - A(t_i))^2.$$

To optimize, we simply need to find the values for when the gradient of E is 0. Taking the partial derivative of E with respect to β_1 and setting it equal to 0 we get the following,

$$0 = E_{\beta_1},$$
$$0 = \frac{2}{N}\sum_{i=1}^{N}(\beta_1 Z_1(t_i) + \beta_2 Z_2(t_i) + \beta_3 Z_3(t_i) + \beta_4 - A(t_i))Z_1(t_i),$$
$$\sum_{i=1}^{N} A(t_i)Z_1(t_i) = \sum_{i=1}^{N}(\beta_1 Z_1^2(t_i) + \beta_2 Z_2(t_i)Z_1(t_i) + \beta_3 Z_3(t_i)Z_1(t_i) + \beta_4 Z_1(t_i)).$$

The rest follow similarly, then we can turn this into a system of linear equations

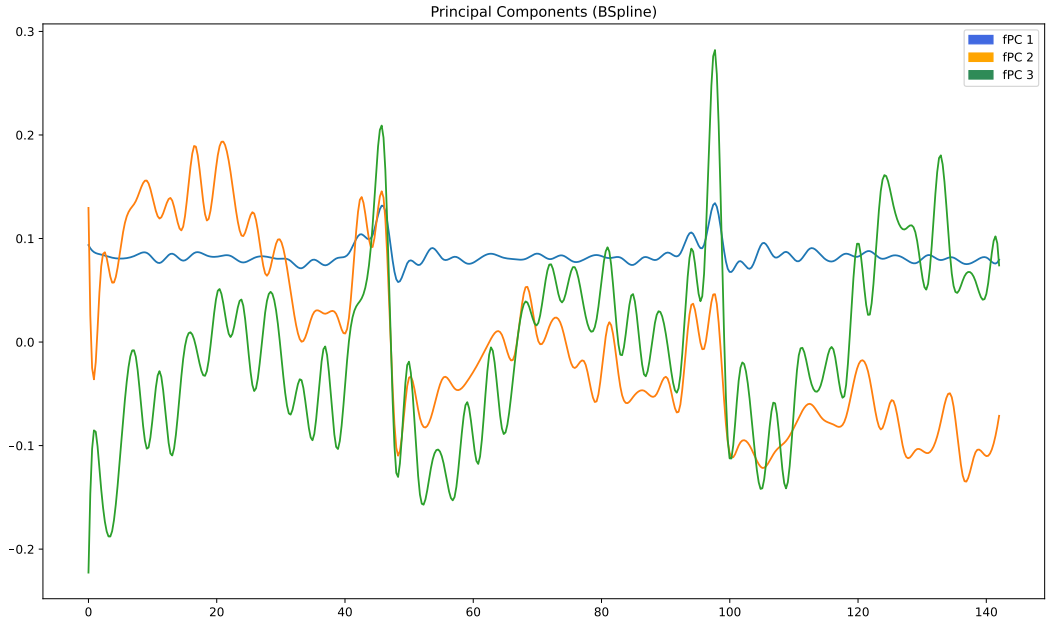

Figure 4.3: The best 3 functional principal components.

and obtain the following,

$$\sum_{i=1}^{N} \begin{bmatrix} Z_1^2(t_i) & Z_2(t_i)Z_1(t_i) & Z_3(t_i)Z_1(t_i) & Z_1(t_i) \\ Z_1(t_i)Z_2(t_i) & Z_2^2(t_i) & Z_3(t_i)Z_2(t_i) & Z_2(t_i) \\ Z_1(t_i)Z_3(t_i) & Z_2(t_i)Z_3(t_i) & Z_3^2(t_i) & Z_3(t_i) \\ Z_1(t_i) & Z_2(t_i) & Z_3(t_i) & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \sum_{i=1}^{N} \begin{bmatrix} A(t_i)Z_1(t_i) \\ A(t_i)Z_2(t_i) \\ A(t_i)Z_3(t_i) \\ A(t_i)Z_4(t_i) \end{bmatrix}.$$

Then with this function Y, we can predict for future moments in time for t_i where $i > N$, as seen in Figure 4.4 where we forecast for the weeks following the 80th week; Figure 4.5 is our prediction with Fourier interpolation.

As seen in Figures 4.4 and 4.5, this regression can predict future trends somewhat well. The average root mean squared error (RMSE) across all the training data was 18174.48 and 19070.76 under Fourier and BSpline interpolation respectively. The reason we are looking at the performance under both types of interpolation is that data may work better in FDA under different types of interpolation.

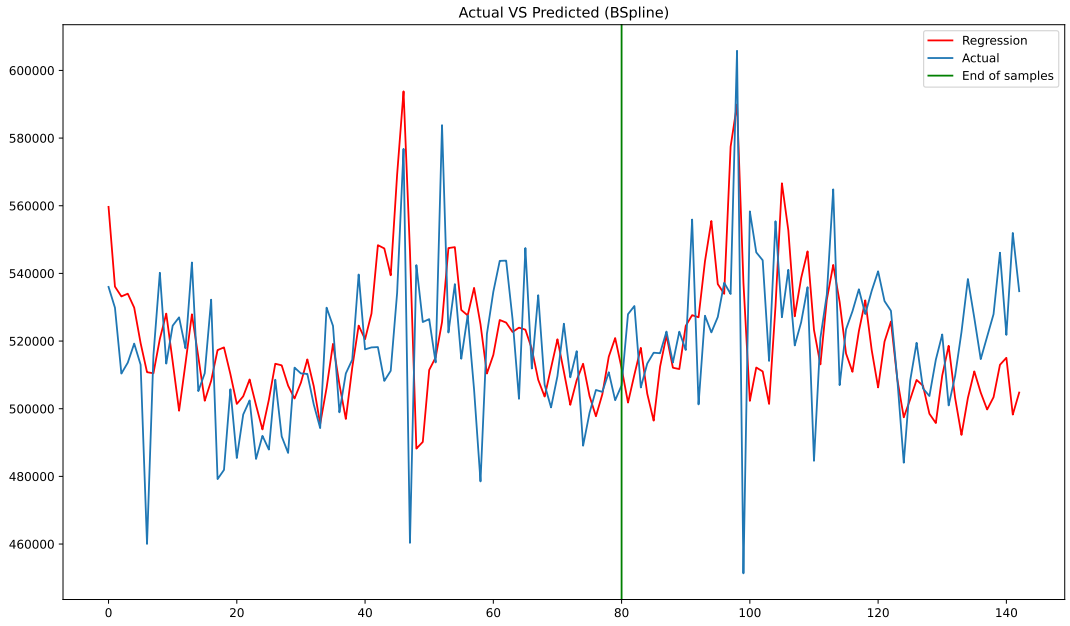

Figure 4.4: Regression plotted against actual store sales under BSpline interpolation.

As one may guess, periodic data is usually better under Fourier interpolation [2]. This guideline lines up with our data, since our data spans across 3 years, and we are able to see periodic trends. Going back to our RMSE though, these errors may seem large, but relative to the average weekly sales for these stores, it translates to about a 2-3 percent error. Running this training and testing split data under traditional PCA and performing regression there, we get an RMSE of 23381.804, which is about a 4-5 percent error. If one wanted to see a bigger difference in performance, in favor of functional PCA, it is best to work with functional data where the amount of time steps is significantly larger than the amount of sample paths [2]. Since, at least for time series data, FDA is performed when the time unit is in minutes or seconds.

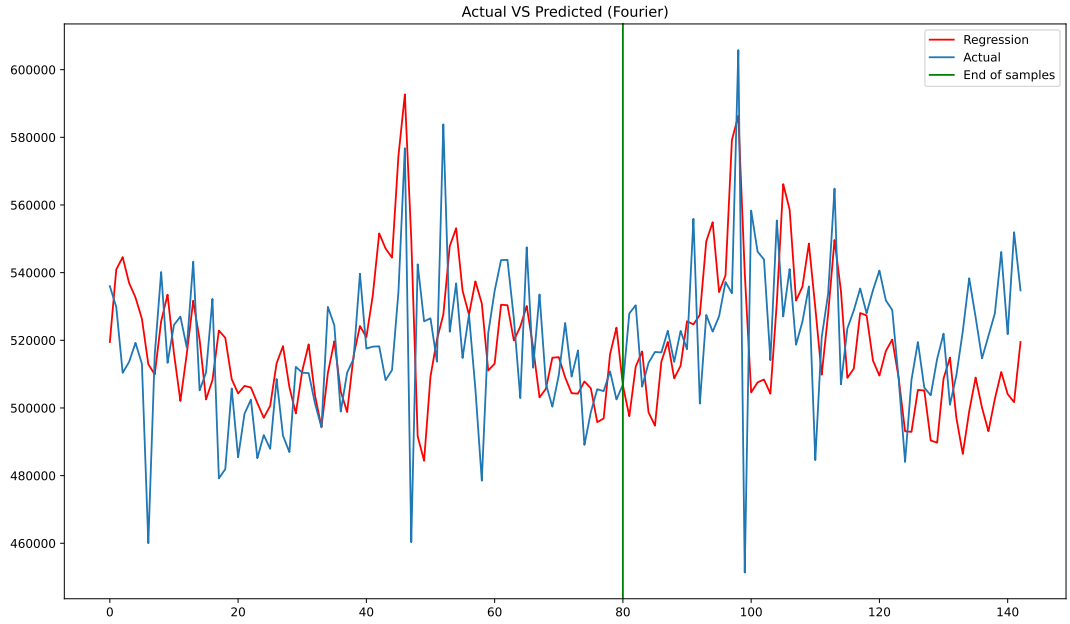

Figure 4.5: Regression plotted against actual store sales under Fourier interpolation.

4.3 Conclusion

Overall, performing regression through functional PCA on functional data is promising in practice as seen by our results, and in general, FDA is promising as a whole. Recall that the extensions do not end here, and one could utilize the cross-covariance operator and perform correlation analysis [2][3], making use of other data features. As we can see, viewing sample paths as curves living in a function space can yield insights one may not observe or expect in performing traditional data analysis. Through this exploration, we can begin to appreciate the richness of FDA from blending statistics and functional analysis.

Bibliography

[1] Billingsley, P. (1999) *Convergence of Probability Measures*, Wiley-Interscience.

[2] Ramsay, J. O. & Silverman, B. W. (2005), *Functional Data Analysis*, Springer.

[3] Eubank, R., & Hsing, T. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators* (Vol. 997). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118762547

[4] Etemadi, N. (1981) *An elementary proof of the strong law of large numbers*. Z. Wahrscheinlichkeitstheorie verw Gebiete 55, 119–122. https://doi.org/10.1007/BF01013465

[5] Baker, C. R. (1973). *Joint Measures and Cross-Covariance Operators*. Transactions of the American Mathematical Society, 186, 273–289. https://doi.org/10.2307/1996566

[6] Baker, C. R., & McKeague, I. W. (1981). *Compact Covariance Operators*. Proceedings of the American Mathematical Society, 83(3), 590–593. https://doi.org/10.2307/2044126

[7] Degras, D. (2017), *Simultaneous confidence bands for the mean of functional data*. WIREs Comput Stat, 9: e1397. https://doi.org/10.1002/wics.1397

[8] Cuevas, A. (2014). *A partial overview of the theory of statistics with functional data*. Journal of Statistical Planning and Inference. 147. 1–23. 10.1016/j.jspi.2013.04.002.

[9] Bickel J. & P, Levina, E. (2004) *Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations.* Bernoulli 10 (6) 989 - 1010, https://doi.org/10.3150/bj/1106314847

[10] Dauxois, J. & Pousse, A. & Romain, Y., (1982). *Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference*, Journal of Multivariate Analysis, Elsevier, vol. 12(1), pages 136-154, March.

[11] Durrett, R. (1996). *Probability: theory and examples*. Belmont, CA: Duxbury Press. ISBN: 0-534-24318-5