AN INTRODUCTION TO THEORETICAL PROPERTIES OF FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

Ngoc Mai Tran
Supervisor: Professor Peter G. Hall

Department of Mathematics and Statistics,
The University of Melbourne.

August 2008

Submitted in partial fulfillment of the requirements of the degree of Bachelor of Science with Honours
Abstract

The term “functional data” refers to data where each observation is a curve, a surface, or a hypersurface, as opposed to a point or a finite-dimensional vector. Functional data are intrinsically infinite dimensional and measurements on the same curve display high correlation, making assumptions of classical multivariate models invalid. An alternative approach, functional principal components analysis (FPCA), is used in this area as an important data analysis tool. However, presently there are very few academic reviews that summarize the known theoretical properties of FPCA. The purpose of this thesis is to provide a summary of some theoretical properties of FPCA when used in functional data exploratory analysis and functional linear regression. Practical issues in implementing FPCA and further topics in functional data analysis are also discussed, however, the emphasis is given to asymptotics and consistency results, their proofs and implications.

Acknowledgements
I would like to thank my teachers, Professors Peter Hall, Konstatin Borovkov, and Richard Huggins, for their guidance and editing assistance.

Ngoc M. Tran, June 2008.
## Contents

Chapter 1  Introduction .............................................. 1  
1.1 The concept of “functional data” ............................. 1  
1.2 Thesis motivation ................................................. 2  
1.3 A historical note .................................................. 3  
1.4 Definitions ......................................................... 4  
1.4.1 FPCA from the kernel viewpoint ......................... 5  
1.4.2 FPCA from linear operators viewpoint ................. 8  
1.4.3 Connections between the two viewpoints ............... 10  
1.4.4 FPCA and PCA ............................................... 11  
1.5 Estimators ......................................................... 13  
1.6 Properties of estimators ....................................... 14  
1.6.1 Properties of the covariance operator ................. 15  
1.6.2 Consistency of the eigenvalue and eigenfunction estimators .... 19  
1.6.3 Asymptotics of the eigenvalue and eigenfunction estimators .... 21  
1.6.4 Eigenspacings and confidence intervals ............... 24  

Chapter 2  Functional linear model ................................. 28  
2.1 Definitions ......................................................... 30  
2.1.1 Some background on inverse problems ................. 31  
2.1.2 Estimators ...................................................... 34  
2.2 Regularization by truncation .................................. 35  
2.2.1 Choice of regularization parameter .................... 36  
2.2.2 Estimators and relation to PCR and FPCA ............ 36  
2.2.3 Choice of m and consistency ............................... 39  
2.3 Mean integrated squared error ................................. 43  
2.3.1 Tikhonov regularization .................................... 44  
2.3.2 MISE of the truncation method .......................... 45  
2.3.3 MISE in Tikhonov regularization ........................ 48  
2.4 Selected topics in FLM ......................................... 49  
2.4.1 Prediction ....................................................... 49  
2.4.2 Central limit theorems ..................................... 52
<table>
<thead>
<tr>
<th>Chapter 3: Practical issues: sampling and smoothing</th>
<th>54</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 From discrete to functional data</td>
<td>54</td>
</tr>
<tr>
<td>3.2 A model for noise</td>
<td>57</td>
</tr>
<tr>
<td>3.3 Sparsity</td>
<td>57</td>
</tr>
<tr>
<td>3.4 Smoothing with a roughness penalty</td>
<td>59</td>
</tr>
<tr>
<td>3.4.1 Ridge regression</td>
<td>59</td>
</tr>
<tr>
<td>3.4.2 A change in norm</td>
<td>60</td>
</tr>
<tr>
<td>Chapter 4: Concluding Remarks</td>
<td>62</td>
</tr>
<tr>
<td>4.1 A summary of the thesis</td>
<td>62</td>
</tr>
<tr>
<td>4.2 Further topics</td>
<td>63</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 The concept of “functional data”

The term “functional data” refers to data where each observation is a curve, a surface, or a hypersurface, as opposed to a point or a finite-dimensional vector. For example, in meteorology, we might want to analyze the temperature fluctuations in Victoria. Then the temperature data collected over time at each station are effectively producing a curve over the observed interval, with, say, 365 measurements made over 365 days of the year. With the emergence of new technology, many datasets contain measurements monitored over time or space, for example: a sensor records the force applied to an object at rate 500 times per second, a computer records position of a digital pen at rate 600 times per second [31]. Although each measurement is discrete and could be viewed as 600 data points rather than one curve, the collection of points possess a certain smoothness property that facilitates the functional data interpretation.

There are several important features of functional data:

- Functional data are intrinsically infinite dimensional.
- Measurements within one curve display high correlation.

These properties render the assumptions made for multivariate analysis incorrect. On the other hand, functional data analysis (FDA) methods exploit smoothness of the data to avoid both the curse of dimensionality and the high correlation between measurements. In particular, the covariance function for functional data is a smooth fixed function, while in high-dimensional multivariate data, the dimension of covariance matrices is assumed to increase with sample size, leading to the “large p, small n” problems [20]. Therefore, in analysis of data arising in various fields such as longitudinal data analysis, chemometrics, econometrics, etc., the functional data analysis viewpoint is becoming increasingly popular, since it offers an alternative way to approach high-dimensional data problems.

An important dimension reduction method for functional data is functional principal component analysis (FPCA), which is the main topic of study in this thesis. In finite dimensional settings, principal components analysis (PCA) was first described
independently by Pearson in 1901 and Hotelling in 1933 as a dimension reduction technique [22]. Finite dimensional PCA relies on expanding the data in the eigenbasis of the covariance matrix. When extended to infinite dimension, this translates to expanding the data in the eigenbasis of the covariance operator. FPCA reduces the functional data to a given number of components such that the representation is optimal in terms of its $L^2$ accuracy. Many of the advantages of PCA in the finite dimensional case carry on to the infinite dimensional context, including:

- It allows analysis of variance-covariance structure, which can be difficult to interpret otherwise.
- As an orthonormal basis expansion method, it gives the optimal basis (in the sense that it minimizes mean $L^2$-error).
- There is a good theoretical support for FPCA.
- The function can be broken down into “principle components”, allowing visualization of the main features of the data.
- The method extends naturally to functional linear regression, an important topic of much practical interests.

There are several approaches to analyzing functional data. In this thesis we concentrate only on FPCA and its applications. Alternative treatments of functional data are briefly discussed in the last chapter. Further references can be found in books such as [31, 13].

### 1.2 Thesis motivation

To date, there are no books or papers summarizing the currently known properties of FPCA. Ramsay and Silverman [31] offer an applied-oriented introduction to the topic, while asymptotics and consistency properties of estimators are scattered over several papers by different authors in different contexts: dimension reduction [12, 17, 18], functional linear modelling [7, 8, 16], time-series analysis [4], longitudinal analysis [36, 20], to name a few. The purpose of this thesis to provide a summary of some known theoretical properties of FPCA when used in functional data exploratory analysis and function linear regression, as well as some interesting theoretical issues which arise from implementing FPCA in practice. The thesis is organized into the following chapters:

Chapter 1 introduces the theory of FPCA from both the linear operator and kernel viewpoints. Asymptotics and consistency of estimators are discussed.

In chapter 2, an application of FPCA to the problem of functional linear model (FLM), or functional linear regression is explored. This can be thought of as a linear
regression model with functional predictor and functional “slope”. This chapter highlights the advantage of the functional view over the multivariate view in the FLM context. The link between FLM and ill-posed inverse problems are analyzed, and the dependence of the estimator’s properties on choice of smoothing parameter is studied.

In practice, the observed “curve” is made up of discrete measurements. Chapter 3 investigates the theoretical impact on FPCA in translating these measurements to the functional form, in particular, the impact on convergence rates of estimators. Models with noise and sparse measurements are considered. The last part of chapter 3 studies some theoretical implications of smoothing with a roughness penalty.

In chapter 4, the concluding chapter, a summary is provided, and further topics in PCA and FDA are briefly mentioned, including generalizations to multivariate FPCA and nonparametric approaches.

1.3 A historical note

On functional data analysis
Ramsay and Silverman [31] noted that functional data have two key features: replication (taking measurements on the same subject repeatedly over different time or different space), and smoothness (the assumption that the underlying curve has a certain degree of smoothness). From this viewpoint, functional data analysis touches base with a few different areas of statistics, and can even be dated back to Gauss in 1809 and Lagrange in 1805 in the problem of estimating comet’s trajectory [31]. Emphasis on analyzing data with replications appeared in texts on ANOVA and longitudinal data analysis in the 1990s. Smoothness is taken into account earlier, motivated by a series of papers on human growth data in 1978 by Largo (the references can be found in [31]). Technical developments, especially in the principal components analysis direction, were motivated by the paper published in 1982 by Dauxois [12]. Subsequently many papers which treated functional principal components analysis under the linear operator viewpoint were published by the French school, with Besse [2], Cardot [7, 10], Ferraty [13], Mas [26, 25] and Bosq [4] being the main contributors (more references can be found in Ramsay and Silverman [31]). From the 1990s onwards, with the development of technology, functional data analysis became a fast growing area of statistics. The texts Functional Data Analysis [31] and Nonparametric Functional Data Analysis [13] are considered to be important texts for this topic.
On functional principal components analysis

In the finite dimensional setting, the first papers on principal components analysis (PCA) were written independently by Pearson in 1901 and Hotelling in 1933. These authors considered PCA as a dimension reduction technique [22]. Eigenanalysis for a symmetric matrix was extended to integral operators with symmetric kernels, leading to the Karhunen-Loève decomposition (Karhunen, 1947, Loève, 1945) [31], which is an important result in the theory of FPCA. The earliest attempt to use PCA for functional data was made in 1958 independently by Rao and Tucker, who applied multivariate PCA without modifications to observed function values [31]. Important asymptotic properties of PCA estimators for the infinite dimensional case were first studied in 1982 by Dauxois [12]. Since then, functional PCA (FPCA) became an important technique in functional data analysis. Many theoretical development in the 1980s and 1990s came from the linear operator viewpoint, as mentioned previously. Practical motivations, and the fact that the audience of these papers largely consist of applied statisticians, led to more recent work which analyze FPCA from the kernel viewpoint, see, for example, Yao et al. [36], Hall et al. [16, 17, 20]. As we shall see, this viewpoint is beneficial in certain applications, for example, the calculation of the kernel operator, or incorporating local smoothing ideas, such as those used in [20, 36]. Therefore, in this thesis we shall introduce FPCA from both perspectives. The two views are interchangable.

1.4 Definitions

Notation

Below we state some notations and definitions which will be used throughout the thesis. Some elementary properties are stated without proofs. Details can be found in texts on linear operators such as Weidmann [34], and in texts on stochastic process such as Bosq [4]. These notations mainly follow that of Bosq [4].

- \((\Omega, \mathcal{A}, P)\): a probability space with sigma algebra \(\mathcal{A}\) and measure \(P\).
- \(L^2(\mathcal{I}) = L^2(\mathcal{I}, \mathcal{B}, \lambda)\): the space of (class of) square-integrable functions on the compact set \(\mathcal{I}\), \(f : \mathcal{I} \rightarrow \mathbb{R}\), \((\int_{\mathcal{I}} f^2)^{1/2} < \infty\). This space is a separable Hilbert space, with inner product \(\langle f, g \rangle := \int_{\mathcal{I}} fg\), and norm \(\|f\| := (\int_{\mathcal{I}} f^2)^{1/2}\).
- \(L^2(\Omega) = L^2(\Omega, \mathcal{A}, P)\): the space of (class of) random variables \(Y : \Omega \rightarrow L^2(\mathcal{I})\) such that \((\int_{\Omega} \|Y\|^2 dP)^{1/2} < \infty\). The function \(Y \mapsto (\int_{\Omega} \|Y\|^2 dP)^{1/2} = \int_{\mathcal{I}} E(Y^2)\) is a norm on \(L^2(\Omega)\).
- \(\mathcal{H}\): a separable Hilbert space. Most of the time, we work with \(\mathcal{H} = L^2(\mathcal{I})\).
• \( \| K \|_\infty \): for \( K \) a linear operator from \( \mathcal{H}_1 \) to \( \mathcal{H}_2 \), this denote the operator norm of \( K \), that is

\[
\| K \|_\infty = \sup_{\psi \in \mathcal{H}_1, \| \psi \|=1} \| K \psi \|.
\]

Operators \( K \) such that \( \| K \|_\infty < \infty \) are called bounded operators. The set of bounded linear operators from \( \mathcal{H}_1 \) to \( \mathcal{H}_2 \) is denoted \( B(\mathcal{H}_1, \mathcal{H}_2) \). Equipped with the \( \| \cdot \|_\infty \) norm, this is a Banach space.

• \( \| K \|_S \): for \( K \in B(\mathcal{H}_1, \mathcal{H}_2) \), this denote the Hilbert-Schmidt norm of \( K \), that is, for any basis \((e_j)_{j=1}^\infty \) of \( \mathcal{H}_1 \)

\[
\| K \|_S = \left( \sum_{j=1}^{\infty} \| K(e_j) \|^2 \right)^{1/2}
\]

This definition is independent of the choice of the basis \((e_j)\). Linear operators \( K \) such that \( \| K \|_S < \infty \) are called Hilbert-Schmidt operators. The set of Hilbert-Schmidt operators from \( \mathcal{H}_1 \) to \( \mathcal{H}_2 \) is denoted \( S(\mathcal{H}_1, \mathcal{H}_2) \). Equipped with the \( \| \cdot \|_S \) norm, this is a Hilbert space.

• \( \mathcal{H}^* \): the dual space of the separable Hilbert space \( \mathcal{H} \). That is, it is the space of continuous linear operators \( T : \mathcal{H} \to \mathbb{R} \). Now, \( \mathcal{H}^* = B(\mathcal{H}, \mathbb{R}) = S(\mathcal{H}, \mathbb{R}) \). We shall use the notation \( \| \cdot \| \) without subscript to denote the usual norm for this space.

• \( x^* \): for \( x \in \mathcal{H} \), \( x^* \) denote its Riesz representation in \( \mathcal{H}^* \).

• \( \| f \| \): for \( f \) an element of \( L^2(\mathcal{I}) \), this denote the norm of \( f \) (with respect to the usual \( L^2 \)-norm).

Assumptions:
For simplicity, in this thesis, we shall work with the Hilbert space \( L^2(\mathcal{I}) \) equipped with the usual \( L^2 \) inner product unless otherwise indicated. Further, we assume that \( \mathcal{I} \subset \mathbb{R} \). The theory of the general case holds many similarities to this case. Some generalizations are mentioned in chapter 4.

1.4.1 FPCA from the kernel viewpoint
Let \( X \) denote a random variable \( X : \Omega \to L^2(\mathcal{I}) \), such that \( X \in L^2(\Omega) \). Equivalently, \( X \) can be seen as a stochastic process defined on a compact set \( \mathcal{I} \), with finite mean-trajectory: \( \int_\mathcal{I} E(X^2) < \infty \). Define \( \mu := E(X) \) to be the mean process of \( X \).
Definition 1. The covariance function of $X$ is defined to be the function $K : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}$, such that

$$K(u, v) := \text{Cov}\{X(u), X(v)\} = E\{[X(u) - \mu(u)][X(v) - \mu(v)]\}.$$ 

Assume that $X$ has continuous and square-integrable covariance function, that is, \(\int \int K(u,v)^2 \, du \, dv < \infty\). Then the function $K$ induces the kernel operator $K : L^2(\mathcal{I}) \rightarrow L^2(\mathcal{I}), \psi \mapsto K\psi$, defined by

$$(K\psi)(u) = \int_{\mathcal{I}} K(u,v)\psi(v) \, dv.$$ 

To reduce notational load, we shall use $K$ to denote the kernel operator. The covariance function $K$ would be referred to as “the (covariance) function $K$”. Usually it should be clear from the context which $K$ is implied.

As noted, FPCA relies on an expansion of the data in terms of the eigenbasis of $K$. The existence of an eigenbasis of $K$ for $L^2(\mathcal{I})$ is guaranteed by Mercer’s lemma, and the expansion of $X$ in this basis is termed the Karhunen-Loève expansion.

Lemma 2. Mercer’s Lemma [4](p24)
Assume that the covariance function $K$ as defined is continuous over $\mathcal{I}^2$. Then there exists an orthonormal sequence $(\psi_j)$ of continuous functions in $L^2(\mathcal{I})$, and a non-increasing sequence $(\kappa_j)$ of positive numbers, such that

$$(K\psi_j)(u) = \kappa_j \psi_j(u), \quad u \in \mathcal{I}, j \in \mathbb{N},$$

and moreover,

$$K(u,v) = \sum_{j=1}^{\infty} \kappa_j \psi_j(u)\psi_j(v), \quad u,v \in \mathcal{I}, \quad (1.1)$$

where the series converges uniformly on $\mathcal{I}^2$. Hence

$$\sum_{j=1}^{\infty} \kappa_j = \int_{\mathcal{I}} K(u,u) \, du < \infty.$$ 

Proof: A proof is given in section 1.4.2, which illustrates the link between Mercer’s lemma and the spectral theorem for compact self-adjoint linear operators. This important fact connects the kernel viewpoint and the linear operator viewpoint.
Under the assumptions and notations of Mercer’s lemma, we have
\[ X(u) = \mu(u) + \sum_{j=1}^{\infty} \sqrt{\kappa_j} \xi_j \psi_j(u), \]  
(1.2)
where \( \xi_j := \frac{1}{\sqrt{\kappa_j}} \int X(v) \psi_j(v) \, dv \) is a random variable with \( E(\xi_j) = 0 \), and
\[ E(\xi_j \xi_k) = \delta_{j,k}, \quad j, k \in \mathbb{N}, \]
where \( \delta_{j,k} \) denote the Kronecker delta. The series (1.2) converges uniformly on \( I \) with respect to the \( L^2(\Omega) \)-norm.

Outline of proof: The key step in the proof is to change the order of integration to obtain:
\[ E\{X(u) \xi_j\} = \kappa_j \psi_j(u). \]
Then we expand \( E\{X(u) - \sum_{j=1}^{n} \sqrt{\kappa_j} \xi_j \psi_j(u)\}^2 \) and use equation (1.1) of Mercer’s lemma to obtain the uniform convergence needed. The expectation of the eigenvalues can be checked via direct computation. □

The key to the Karhunen-Loève expansion, and is also one of the advantages of the FPCA approach, is that \( X \) is broken up into orthogonal components with uncorrelated coefficients. The eigenvalue \( \kappa_j \) can be interpreted as a measure of the variation in \( X \) in the \( \psi_j \) direction. The idea of FPCA is to retain the first \( M \) terms in the Karhunen-Loève expansion as an approximation to \( X \)
\[ \hat{X} = \sum_{j=1}^{M} \sqrt{\kappa_j} \xi_j \psi_j, \]
and hence achieve dimension reduction. This can be seen as projecting \( X \) onto an \( M \)-dimensional space, spans by the first \( M \) eigenfunctions of \( K \), which are those with the largest eigenvalues \( \kappa_j \). This projection could be done with other basis, but the eigenbasis \( (\psi_j) \) is “optimal” in the sense that it minimizes the expected error in the M-term approximation. This lemma is stated and proved in section 1.4.3.

In the FPCA literature, there is no standard notation for the eigenvalues, eigenvectors or the covariance function. Some authors denote the covariance operator by \( \Gamma \), the eigenvalues as \( \lambda_j \), and they tend to work with the eigenprojections, which are projections onto the space spans by the eigenfunctions, instead of the actual eigenfunctions. Authors who approach FPCA from a kernel viewpoint such as Hall et al. [17], Yao et al. [36], tend to call the covariance operator \( K \), the eigenfunctions \( \psi_j \), and the eigenvalues either \( \kappa_j \) or \( \lambda_j \). Since the kernel viewpoint was introduced first in this thesis, our notations follow those of Hall [17].

7
For the word ‘principal component’ itself, some authors use this in reference to $\psi_j$, others use it to refer to $\xi_j$ [22]. Here we shall follow the definitions used in Dauxois [12].

**Definition 4.** Let $(\xi_j), (\psi_j), (\kappa_j)$ be defined as in the Karhunen-Loève theorem. Let $\Pi_j$ be the projection of $L^2(I)$ onto the space spans by eigenfunctions with corresponding eigenvalue $\kappa_j$. Then the $(\Pi_j)$ are called eigenprojections. The $(\psi_j)$ are called the principal factors of $X$, the $(\kappa_j)$ are called the principal values, and the $(\xi_j)$ are called the standardized principal components of $X$, and the values $(\langle X, \psi_j \rangle) = (\sqrt{\kappa_j} \xi_j)$ are called the (unstandardized) principal components of $X$.

Now, we shall introduce the linear operator viewpoint and study the properties of the estimators.

### 1.4.2 FPCA from linear operators viewpoint

**Definition 5.** The covariance operator of $X$ is the map $C_X : (L^2(I))^* \to L^2(I)$, defined by

$$C_X(x^*) = \mathbb{E}[x^*(X - \mu)(X - \mu)], \quad x^* \in (L^2(I))^*.$$  

(Recall that $(L^2(I))^*$ stands for the space of continuous linear operators from $L^2(I)$ to $\mathbb{R}$).

This definition applies to general Banach spaces. Since $L^2(I)$ is a Hilbert space, by the Riesz representation theorem, the covariance operator can be viewed as

$$C_X : L^2(I) \to L^2(I), \quad C_X(x) = \mathbb{E}[(X - \mu, x)(X - \mu)], \quad x \in L^2(I).$$

With some computation, we recognize that this is the kernel operator $K$ defined in definition 1. Indeed, let $x \in L^2(I)$. Then $C_X : L^2(I) \to L^2(I)$, with

$$C_X(x)(u) = \mathbb{E}[(X - \mu, x) (X - \mu)(u)]$$

$$= \int_{\Omega} \int_{I} (X - \mu)(v, \omega)x(v) dv (X - \mu)(u, \omega)dP(\omega)$$

$$= \int_{I} \mathbb{E}[(X - \mu)(u)(X - \mu)(v)]x(v) dv$$

$$= \int_{I} K(u, v)x(v) dv = Kx(u).$$
Therefore, $K = C_X$. From now on, we shall continue to use $K$ for the covariance operator, in accordance with the previous section.

The following lemma states some properties of the covariance operator $K$. These properties parallel those of a covariance matrix. The lemma suggests that PCA in the functional setting relies on the same principles as in the multivariate setting.

**Lemma 6.** The covariance operator $K$ is a linear, self-adjoint, positive semidefinite operator.

**Proof:** By direct computation, using the kernel operator definition. The first property is trivial. The second property relies on the symmetry of the covariance function: $K(u, v) = K(v, u)$. The third property relies on changing the order of integration to obtain $\langle K\psi, \psi \rangle = E|\langle X, \psi \rangle|^2$. \qed

An important result for compact self-adjoint operators is the spectral theorem, which bear close resemblance to Mercer’s lemma and the Karhunen-Loève expansion. The proof of this classical result is omitted, and can be found in [34] (p166).

**Theorem 7. Spectral theorem for compact self-adjoint linear operators** (abbr.: Spectral theorem) [34](p166)

Let $T$ be a compact self-adjoint bounded linear operator on a separable Hilbert space $\mathcal{H}$. Then there exists a sequence of real eigenvalues of $T$: $|\kappa_1| \geq |\kappa_2| \geq \cdots \geq 0$, $\kappa_j \to 0$ as $j \to \infty$. Furthermore, let $\Pi_j$ be the orthogonal projections onto the corresponding eigenspaces. Then the $\langle \Pi_j \rangle$ are of finite rank, and

$$T = \sum_{j=1}^{\infty} \kappa_j \Pi_j,$$

where the series converges in the operator norm.

**Corollary 8.** There exist eigenfunctions $(\psi_j)$ of $T$ corresponding to the eigenvalues $(\kappa_j)$, which form an orthonormal basis for $\mathcal{H}$. Further, if $T$ is positive semidefinite, then $\kappa_j \geq 0$ for all $j$.

**Proof:** Since $\mathcal{H}$ is separable, if we choose a (possibly countably infinite) orthonormal basis for the kernel space of $T$, and amalgamate them with the $(\psi_j)$ in the
theorem to create one sequence, then this is an orthonormal basis for $\mathcal{H}$.\footnote{We shall denote this amalgated sequence by $(\psi_j)$, since the newly amalgated functions are also eigenfunctions of $T$ with eigenvalue 0.} The positive semidefiniteness comes from direct computation.

\section*{1.4.3 Connections between the two viewpoints}
Authors who approach FPCA from the kernel viewpoint tend to work with the covariance function, while those who use the operator viewpoint often work with the covariance operator. Using the notation of tensor product spaces (see Weidmann [34], p37), the covariance function $K$ can be written as

$$
covariance\ function\ K = E\{(X - \mu) \otimes (X - \mu)\},
$$

and the covariance operator can be written as

$$
C_X = operator\ K = E\{(X - \mu)^* \otimes (X - \mu)\}.
$$

The duality is clear: the covariance function and covariance operator are duals via the Riesz representation theorem. Therefore, the two views are interchangeable, which implies that results on the operator $K$ can be easily given an interpretation in terms of the covariance function $K$. Further, convergence of the covariance function (for example, asymptotic normality) implies convergence of the covariance operator, and vice versa.

There is a strong connection between the spectral theorem, Mercer’s lemma and Karhunen-Loève expansion. By the spectral theorem, (the completion of) $(\psi_j)$ is an orthonormal basis for $L^2(\mathcal{I})$. So for any $x \in L^2(\mathcal{I})$, Plancherel’s theorem [34] (p38) gives

$$
x = \sum_{j=1}^{\infty} \psi_j \langle x, \psi_j \rangle = \sum_{j=1}^{\infty} \psi_j \int_{\mathcal{I}} x(v) \psi_j (v) \, dv,
$$

where the series converges in the $L^2$-norm. This implies convergence in the $L^\infty$-norm, which implies uniform convergence of $\sum_{j=1}^{\infty} \psi_j (u) \int_{\mathcal{I}} x(v) \psi_j (v) \, dv$ on $\mathcal{I}$. The Karhunen-Loève theorem extends this expansion to the random variable $X$ itself. The uniform convergence part relies on Mercer’s lemma, which can be proved using the spectral theorem.

**Proof of Mercer’s lemma based on the spectral theorem**
From lemma 6, we only need to show compactness to apply the spectral theorem.
on $K$. This will be the case if we can show that $K$ is a Hilbert-Schmidt operator. Indeed: let $(e_i)$ be an orthonormal basis for $\mathcal{H} := L^2(I)$. Then $(e_i \otimes e_j)$ is an orthonormal basis in $L^2(I^2)$, and

$$
\int_{I^2} K^2(u, v) \, du \, dv = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \left( \int_{I^2} K(u, v) g_j(v) g_i(u) \, du \, dv \right)^2
$$

$$
= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \left( \int_{I} K g_j(u) g_i(u) \, du \right)^2
$$

$$
= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \langle K g_j, g_i \rangle = \sum_{j=1}^{\infty} \| K g_j \|^2 < \infty.
$$

So $K$ is a Hilbert-Schmidt operator, hence is compact. Apply the spectral theorem to $K$, we obtain an orthonormal basis $(\psi_j)$ for $L^2(I)$, with

$$(K \psi_j)(u) = \kappa_j \psi_j(u), \quad u \in I, j \in \mathbb{N}.$$

Exploit the continuity of the inner product, and the fact that $\langle x, y \rangle = 0$ for all $y \in L^2(I)$ implies $x = 0$, we can prove that

$$K(u, v) = \sum_{j=1}^{\infty} \kappa_j \psi_j(u) \psi_j(v),$$

where the convergence is uniform on $I^2$. Therefore

$$\sum_{j=0}^{\infty} \kappa_j = \int_{I} K(u, u) \, du < \infty.$$

This proves Mercer’s lemma. The proof highlights the relationship between the lemma and the spectral theorem, and hence confirms the connection between the kernel viewpoint and the linear operator viewpoint.

### 1.4.4 FPCA and PCA

It is worth mentioning the parallel and distinction between FPCA and PCA. In the later case, the vector $X$ is decomposed into components using the spectral theorem for symmetric real matrices [22] (p13). Since symmetric real matrices are automatically compact, this theorem is a special case of the spectral theorem stated above. Many properties of PCA carry over to the infinite dimensional settings, for example, “the best $M$-term approximation” property, noted in [12, 31].
Lemma 9. For any fixed $M \in \mathbb{N}$, the first $M$ principal factors of $X$ satisfy

\[
(\psi_i) = \arg\min_{(\phi_i) : \langle \phi_i, \phi_j \rangle = \delta_{ij}} E \left\| X - \sum_{i=1}^{M} \langle X, \phi_i \rangle \phi_i \right\|^2,
\]  

(1.3)

where $\delta_{ij}$ is the Kronecker’s delta.

Proof: The proof is similar to that of the finite dimensional setting. Let $(\phi_i)$ be an orthonormal basis of $L^2(I)$ which give the expansion

\[
X = \sum_{i=1}^{\infty} \langle X, \phi_i \rangle \phi_i.
\]

By orthonormality of the $(\phi_i)$, the minimization problem in the right-hand side of equation (1.3) is equivalent to maximizing $E \left( \sum_{i=1}^{M} \langle X, \phi_i \rangle^2 \right)$ subjecting to $\|\phi_i\| = 1$ for each $i$.\(^2\) This can be solved by induction. Consider the case $M = 1$. Under the constraint $\|\phi\| = 1$, we want to maximize

\[
E \langle X, \phi \rangle^2 = E \left( \int_I (X - \mu)\phi(u) du \int_I (X - \mu)(v)\phi(v) dv \right)
= E \left( \int_I \int_I (X - \mu)(u)(X - \mu)(v)\phi(v)\phi(u) dv du \right)
= \int_I \int_I K(u, v)\phi(v)\phi(u) dv du
= \langle K\phi, \phi \rangle.
\]

Subjecting to $\|\phi\| = 1$, $\langle K\phi, \phi \rangle$ is maximized when $\phi$ is the eigenfunction of $K$ with the largest eigenvalue, that is, $\phi = \psi_1$. An induction argument on $M$ shows that the basis needed consists of eigenfunctions of $K$ arranged in a non-increasing order of corresponding eigenvalues, which is the FPCA basis $(\psi_i)$.

Much of the theory on the estimators of FPCA are also similar to the finite dimensional case, such as the parametric $O(n^{-1/2})$ convergence rates for the covariance operator, eigenvalues and eigenfunctions, asymptotic normality, and the eigenspacings problem. The distinction between the two cases only becomes clear when compactness plays an important role. We shall analyze this in the functional linear context regression in chapter 2.

\(^2\)For this reason, the principal components basis is also sometimes derived as the orthonormal basis which maximizes the variance in the reduced data.
1.5 Estimators

Suppose we are given a set \( X = \{X_1, \ldots, X_n\} \) of independent random functions, identically distributed as \( X \). Define \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \). We are interested in estimating the eigenvalues \( (\kappa_j) \) and the principal factors \( (\psi_j) \) of \( X \) from this sample. The principal components can be obtained by numerical evaluation of the integral \( \int_{\mathcal{I}} X(u)\psi_j(u) \, du \).

**Definition 10.** Let \( \bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i \) denote the sample mean. The empirical approximation of \( K(u, v) \) is

\[
\hat{K}(u, v) = \frac{1}{n} \sum_{i=1}^{n} [X_i(u) - \bar{X}(u)][X_i(v) - \bar{X}(v)].
\]

This induces an integral operator, which is called the **sample covariance operator**

\[
\hat{K} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^* \otimes (X_i - \bar{X}).
\]

It is clear that \( \hat{K} \) is also a self-adjoint, compact, positive semidefinite linear operator. So by Mercer’s lemma, there exists an orthonormal basis \( (\hat{\psi}_j) \) for \( L^2(\mathcal{I}) \) consisting of eigenfunctions of \( \hat{K} \), with eigenvalues \( (\hat{\kappa}_j) \), such that

\[
\hat{K}(u, v) = \sum_{j=1}^{\infty} \hat{\kappa}_j \hat{\psi}_j(u) \hat{\psi}_j(v).
\]

Denote the dimension of the range of \( \hat{K} \) by \( r \). Since \( \hat{K} \) maps onto the space spans by the \( n \) functions \( (X_i - \bar{X})^*_{i=1} \) of \( L^2(\mathcal{I}) \), we have \( r \leq n \). Order the eigenvalues so that \( \hat{\kappa}_1 \geq \hat{\kappa}_2 \geq \cdots \geq 0 \), with \( \hat{\kappa}_j = 0 \) for \( j \geq r + 1 \). Apply the spectral theorem, we can write

\[
\hat{K} = \sum_{j=1}^{r} \hat{\kappa}_j \hat{\Pi}_j,
\]

where \( \hat{\Pi}_j \) corresponds to the projection onto the eigenspace with eigenvalue \( \hat{\kappa}_j \).

Complete the sequence \( (\hat{\psi}_j) \) to obtain a complete orthonormal basis for \( L^2(\mathcal{I}) \), then for each \( X_i \)

\[
X_i = \sum_{j=1}^{\infty} \langle X_i, \hat{\psi}_j \rangle \hat{\psi}_j.
\]
In most statistical studies, such as those quoted in the reference, \( \hat{\psi}_j \) and \( \hat{\kappa}_j \) are treated as estimators for \( \psi_j \) and \( \kappa_j \). To make this concrete, we shall state this as a definition.

**Definition 11.** In this thesis, unless otherwise indicated, \( \hat{\psi}_j \) is the **eigenfunction estimator** for \( \psi_j \), \( \hat{\kappa}_j \) is the **eigenvalue estimator** for \( \kappa_j \), and \( \hat{\Pi}_j \) is the **eigenprojection estimator** for \( \Pi_j \).

**Existence of the estimators**

The eigenvalues are well-defined, hence so are the associated eigenprojections. The eigenfunctions are slightly more complicated. First, since the eigenfunctions \( \psi_j \) are unique up to a change of sign, without loss of generality, we can assume that \( \hat{\psi}_j \) is chosen such that \( \langle \psi_j, \hat{\psi}_j \rangle \geq 0 \), to ensure that \( \hat{\psi}_j \) can be viewed as an estimator of \( \psi_j \) and not \( -\psi_j \). Second, note that if an eigenvalue has multiplicity \( m > 1 \), then the associated eigenspace is of dimension \( m > 1 \). Hence there is no unique eigenfunction basis for this space: rotation of a basis would lead to new ones. Therefore, in this thesis, for simplicity we shall primarily work with the case where eigenvalues are distinct, that is, \( \kappa_1 > \kappa_2 > \ldots \geq 0 \).

Given the two conditions above, the eigenfunctions are well-defined. However, in terms of estimations, closely spaced eigenvalues can cause problems. Specifically, if \( \kappa_j - \kappa_{j+1} = \epsilon \) for some small \( \epsilon > 0 \), the eigenfunctions can be so close that it is difficult to estimate. Indeed, we have

\[
K(\psi_j - \psi_{j+1}) = \kappa_{j+1}(\psi_j - \psi_{j+1}) + \epsilon \psi_j
\]

If \( \epsilon \) is very small, then the function \( \psi_j - \psi_{j+1} \) is only \( \epsilon \) away from being an eigenfunction of \( K \). Intuitively, we need large \( n \) to distinguish the two eigenfunctions \( \psi_j \) and \( \psi_{j+1} \), and this leads to slow convergence rate of the estimators \( \hat{\psi}_j \) and \( \hat{\psi}_{j+1} \). In the finite dimensional context, this problem is known as the **eigenspacings problem**, and the effect is known as **instability of the principal factors** [22]. We shall discuss this in details in the next section.

### 1.6 Properties of estimators

It is natural to raise the issue of the consistency and asymptotic properties of the estimators \( \hat{K}, \hat{\psi}_j \) and \( \hat{\kappa}_j \) defined above. From a practical viewpoint, these results would allow us to derive confidence intervals and hypothesis tests. Under mild assumptions, consistency results was derived by Dauxois et al. [12], using the method
of spectral perturbation. More specific cases, such as convergence of covariance operators in linear processes, were studied by Mas [25]. Many models in time series analysis, such as continuum versions of moving average and autoregressive processes are of this type. The extra assumption on the covariance of the data (for example, autoregressive) allows methods of operator theory to be applied effectively in this context. Bosq [4] gives an extensive study of consistency, asymptotics and convergence rates of FPCA estimators in autoregressive processes.

In this section we shall start with some simple but important bounds that leads to consistency theorems. Asymptotic properties of the estimators will be summarized, and the implications of these theorems will be analyzed. In particular, emphasis will be given to the effects of eigenvalues spacing on convergence rates. This effect will also be seen in the second chapter, when we look at functional linear models.

### 1.6.1 Properties of the covariance operator

**The effect of estimating \( \mu \)**

In many papers in the literature, such as the work of Dauxois [12], Bosq [4], Cardot [7], it is often assumed that \( E(X) = \mu = 0 \), and that we observe \((X_1, \ldots, X_n)\) random variables, independent, identically distributed as \( X \). This is equivalent to approximating \( K(u, v) \) by the function

\[
\tilde{K}(u, v) = \frac{1}{n} \sum_{i=1}^{n} [X_i(u) - \mu(u)][X_i(v) - \mu(v)],
\]

which induces an integral operator

\[
\tilde{K} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^* \otimes (X_i - \mu).
\]

From there, the authors deduced statements on the covariance operator \( \tilde{K} \) and its eigenvalues and eigenfunctions (see, for example, Dauxois [12]).

Since we shall consider higher order terms in the theorem by Hall et al. [17], we shall work with \( \hat{K} \) instead of \( \tilde{K} \). It is sometimes more convenient to prove bounds for \( \tilde{K} \) by proving it for \( \hat{K} \) first. The following lemma ensures that asymptotic properties of \( \hat{K} \) up to order \( O(n^{-1/2}) \) are the same as those of \( \tilde{K} \).

**Lemma 12.** Assume that \( E\|X\|^4 < \infty \). The term that comes from using \( \bar{X} \) instead of \( \mu \) is negligible for asymptotic results of order \( O(n^{-1/2}) \). Specifically, we...
mean

\[ E\|\hat{K} - \tilde{K}\|^2_S = O(n^{-2}), \]

which, by Jensen’s inequality, implies

\[ E\|\hat{K} - \tilde{K}\|_S = O(n^{-1}), \]

**Proof:** Consider

\[ E\|\hat{K} - \tilde{K}\|^2_S = E\|(\bar{X} - \mu)^* \otimes (\bar{X} - \mu)\|^2_S = E\|\bar{X} - \mu\|^4 \]

where the norm on \( \bar{X} - \mu \) is the \( L^2 \)-norm in \( L^2(I) \). Let \( Y_i := X_i - \mu \), then the \( Y_i \) are independent, identically distributed as \( Y = X - \mu \), with \( E(Y) = 0 \). Then, consider

\[
\left( \sum_{i=1}^n \|Y_i\|^2 \right)^2 = \left( \sum_i \sum_{j \neq i} \langle Y_i, Y_j \rangle + \sum_i \|Y_i\|^2 \right)^2 = \left( \sum_i \|Y_i\|^2 \right)^2 + 2 \sum_i \|Y_i\|^2 \sum_i \sum_{j \neq i} \langle Y_i, Y_j \rangle + \left( \sum_i \sum_{j \neq i} \langle Y_i, Y_j \rangle \right)^2.
\]

Taking expectation, we see that

\[ E \left( \sum_i \|Y_i\|^2 \right)^2 = nE\|Y\|^4 + (n^2 - n)(E\|Y\|^2)^2. \]

By independence of the \( (Y_i) \), we have

\[ E \left( 2 \sum_i \|Y_i\|^2 \sum_i \sum_{j \neq i} \langle Y_i, Y_j \rangle \right) = 0. \]

Again, utilizing independence of the \( Y_i \) and Cauchy-Schwarz inequality, we obtain

\[ E \left( \sum_i \sum_{j \neq i} \langle Y_i, Y_j \rangle \right)^2 = 2 \sum_i \sum_{j \neq i} E \langle Y_i, Y_j \rangle^2 \leq 2(n^2 - n)E(\|Y\|^2)^2; \]

This gives

\[ E\|\bar{X} - \mu\|^4 = \frac{1}{n^4} E \left( \|\sum_{i=1}^n Y_i\|^2 \right)^2 = O(n^{-2}). \]
Now we state some important results for the estimator $\hat{K}$. By direct computation, we can see that $\frac{n}{n-1}E\hat{K}(u, v) = K(u, v)$. So $\hat{K}$ is asymptotically unbiased.\(^3\) An application of the strong law of large numbers guarantees almost sure convergence for $\hat{K}$:

Lemma 13. Dauxois et al. [12]

Rewrite $\hat{K}$ as $\hat{K}_n$ to emphasize the dependency of $\hat{K}$ on $n$. Then the sequence of random operators $(\hat{K}_n)_{n\in\mathbb{N}}$ converges almost surely to $K$ in $\mathcal{S} = \mathcal{S}(L^2(I), L^2(I))$.

Proof: Consider the unbiased version of $\hat{K}$, that is,

$$\bar{K} := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^* \otimes (X_i - \bar{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i^* \otimes X_i - \frac{1}{n(n-1)} \sum_{i} \sum_{j \neq i} X_i^* \otimes X_j.$$  

Since the $(X_i)$ are independent, $(X_i^* \otimes X_i)$ are i.i.d integrable random variable from $(\Omega, \mathcal{A}, P)$. Therefore, by the strong law of large numbers (SLLN) in the separable Hilbert space $\mathcal{S}$, $\frac{1}{n} \sum_{i=1}^{n} X_i^* \otimes X_i \overset{a.s.}{\to} E(X^* \otimes X)$. To apply strong law of large numbers on the second term, note that for $j \neq i$, $X_i$ and $X_j$ are independent. Therefore, for each fixed $X_i^*$, the SLLN applies for the sequence $\frac{1}{n-1} \sum_{j \neq i} X_i^* \otimes X_j$ in the separable Hilbert space $X_i^* \otimes L^2(I)$, therefore

$$\frac{1}{n-1} \sum_{j \neq i} X_i^* \otimes X_j \overset{a.s.}{\to} X_i^* \otimes \mu.$$  

Then apply the SLLN again for $\frac{1}{n} \sum_{i=1}^{n} X_i^* \otimes \mu$ in the space $L^2(I)^* \otimes \mu$ to obtain

$$\frac{1}{n} \sum_{i=1}^{n} X_i^* \otimes \mu \overset{a.s.}{\to} \mu^* \otimes \mu.$$  

A delta-epsilon argument can be set up to show that this indeed gives almost sure convergence for the term $\frac{1}{n(n-1)} \sum_{i} \sum_{j \neq i} X_i^* X_j$ to $\mu^* \otimes \mu$. Combining this property with the result for the first term, the lemma is proved.\(^\square\)

\(^3\)Unlike the finite dimensional case, the covariance estimator in FPCA is often taken to be the biased term. This partly stems from the fact that many papers in literature, for example, [12, 4, 7] tend to ignore the effect of estimating $\mu$. Ramsay and Silverman [31] (chapter 9) noted that the correction term does not make too much differences in practice. Therefore, we shall work with the biased estimator $\hat{K}$, and only refer to the unbiased version to simplify proofs.
Since the norm \( \| \cdot \|_S \) is stronger than the norm \( \| \cdot \|_\infty \), this also implies that \( \hat{K}_n \to K \) almost surely in the uniform norm of \( S \). Under the only assumption that the covariance function of \( X \otimes X \) is bounded, we obtain \( O(n^{-1/2}) \) convergence for the mean.

**Lemma 14.** (Cardot et al.) [7]
If \( E\|X\|^4 < \infty \), then
\[
E \left\| \hat{K} - K \right\|_\infty^2 \leq \frac{E \|X - \mu\|^4}{n}.
\]
Therefore,
\[
E\|\hat{K} - K\|^2_\infty \leq \frac{2E \|X - \mu\|^4}{n} + O(n^{-2}).
\]

**Proof:** The proof for the inequality with \( \hat{K} \) is a direct computation, details can be found in Cardot [7]. To obtain the bound for \( \hat{K} \), note that
\[
\|\hat{K} - K\|^2_\infty \leq 2\|\hat{K} - \tilde{K}\|^2_\infty + 2\|\tilde{K} - K\|^2_\infty,
\]
and
\[
2\|\hat{K} - \tilde{K}\|^2_\infty \leq 2\|\hat{K} - K\|^2_3.
\]
Then, by lemma 12 \( E\|\hat{K} - \tilde{K}\|_3^2 = O(n^{-2}) \). The result follows.

Apply the central limit theorem for sequence of independent identically distributed (i.i.d) random variables valued in a Hilbert space (see, for example, [4]), we obtain an asymptotic result for the covariance function \( \hat{K} \).

**Lemma 15.** (Dauxois et al.) [12]
Let \( Z := n^{1/2}(\hat{K} - K) \) denote the scaled difference between the covariance functions \( \hat{K} \) and \( K \). Then the function \( Z \) converges in distribution to a Gaussian process, say, \( Z \), with mean 0 and covariance function \( K_Z \) that of the random variable \( (X - \mu) \otimes (X - \mu) \). Explicitly,
\[
K_Z = \sum_{i,j,k,l \in \mathbb{N}} \sqrt{\kappa_i \kappa_j \kappa_k \kappa_l} E(\xi_i \xi_\ell \xi_j \xi_k)(\psi_i \otimes \psi_j) \otimes (\psi_k \otimes \psi_l)
- \sum_{i,j \in \mathbb{N}} \kappa_i \kappa_j (\psi_i \otimes \psi_j) \otimes (\psi_j \otimes \psi_j)
\]

**Proof:** Write
\[
n^{1/2}(\hat{K} - K) = n^{1/2}(\hat{K} - \tilde{K}) + n^{1/2}(\tilde{K} - K).
\]

18
The key to this proof lies in lemma 12 and the central limit theorem (CLT) for a sequence of independent identically distributed random variables (iid RV) valued in a Hilbert space (see, for example, [4]). To set up the machinery, take the tensor product of the Hilbert space \( L^2(\mathcal{I}) \) with itself to obtain \( L^2(\mathcal{I}^2) \). This space is isomorphic to \( L^2(\mathcal{I}^2) \). Let \( (X - \mu) \otimes (X - \mu) : \Omega \rightarrow L^2(\mathcal{I}^2) \) denote the random variable mapping to \( L^2(\mathcal{I}^2) \), whose images are functions of the form \((u, v) \mapsto (X - \mu) \otimes (X - \mu)(u, v) = (X - \mu)(u)(X - \mu)(v)\). Note that the covariance function \( K \) can be written as

\[
K = E\{(X - \mu) \otimes (X - \mu)\}.
\]

and the estimator \( \hat{K} \) can be written as

\[
\hat{K} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu) \otimes (X_i - \mu).
\]

Now, the space \( L^2(\mathcal{I}^2) \) is also a Hilbert space. Then, by the CLT for a sequence of iid RV mapping to a Hilbert space,

\[
n^{1/2}(\hat{K} - K) = n^{-1/2} \sum_{i=1}^{n} (X_i \otimes X_i - E(X \otimes X)) \rightarrow \mathcal{N}(0, C)
\]

where \( C \) is the covariance function of \((X - \mu) \otimes (X - \mu)\), which is \( E\{(X - \mu) \otimes (X - \mu) \otimes (X - \mu) \otimes (X - \mu)\}\).

Lemma 12 gives

\[
E\|\hat{K} - \hat{K}\|^2 = O(n^{-2})
\]

hence

\[
E\|n^{1/2}(\hat{K} - \hat{K})\|^2 = O(n^{-1}),
\]

that is, \( n^{1/2}(\hat{K} - \hat{K}) \) converges in mean-squared to 0, which implies that it converges to 0 in distribution. Hence the term \( n^{1/2}(\hat{K} - \hat{K}) \) makes a negligible contribution in convergence in distribution of \( Z \). The result follows.

1.6.2 Consistency of the eigenvalue and eigenfunction estimators

In one of the first papers on the theoretical properties of FPCA, Dauxois et al. [12] used the methods of spectral perturbation to analyzed the properties of \( \hat{\kappa}_j, \hat{\Pi}_j \) and \( \hat{\psi}_j \). The idea is this: if we have a linear operator \( T \), and another linear operator \( \hat{T} \) which is a small deviation from \( T \), we want to know how much does the spectrum of \( \hat{T} \) deviate from that of \( T \). This is known as the spectral perturbation problem [24], which is one of the areas of perturbation theory for linear operators, and is
a large and developed field of mathematics. Therefore, in this section, we shall not give details of Dauxois’ results, but only attempt to mention the geometric interpretation of the proof where appropriate, and concentrate on analyzing the statistical implications.

Firstly, we consider an important result which gives a bound on the eigenvalues and eigenfunctions in terms of the operator norm of the covariance operator. Versions of this result appeared in Dauxois et al. [12] and Bosq [4].

**Proposition 16.** (Bosq) [4], lemma 4.1, 4.3

\[
\sup_{j \geq 1} |\hat{\kappa}_j - \kappa_j| \leq \| \hat{K} - K \|_{\infty}.
\]

If the eigenvalues are all distinct (that is, \(\kappa_1 > \kappa_2 > \cdots\)), then

\[
\left\| \hat{\psi}_j - \psi_j \right\| \leq \frac{2\sqrt{2}}{\rho_j} \left\| \hat{K} - K \right\|_{\infty},
\]

where \(\rho_j := \min(\kappa_j - \kappa_{j+1}, \kappa_{j-1} - \kappa_j)\) if \(j \geq 2\), and \(\rho_1 = \kappa_1 - \kappa_2\).

**Proof:** The results come from direct computation. As a consequence of theorem 7,

\[
\kappa_j = \|K - \sum_{i=1}^{j-1} \Pi_i\|_{\infty} = \min_{l \in \mathcal{L}_{j-1}} \|K - l\|_{\infty},
\]

where \(\mathcal{L}_{j-1}\) is the space of bounded operators from \(L^2(I)\) to \(L^2(I)\) such that the range of \(l\) is of dimension \(j - 1\) or less. A similar identity holds for \(\hat{\kappa}_j\). The result for the eigenvalues follows by the triangle inequality. For the eigenfunctions, we expand \(K\hat{\psi}_j - \kappa_j\hat{\psi}_j\) and \(\hat{\psi}_j - \psi_j\) in terms of the \((\psi_i)_{i \in \mathbb{N}}\) basis, and use the bound on the eigenvalues to obtain the result needed. Details of both proofs can be found in Bosq [4], lemma 4.1, 4.3.

Proposition 16 has several important implications. Firstly, coupled with lemma 13, it is clear that the eigenvalue estimators \(\hat{\kappa}_j\) converges almost surely and uniformly to their true values. This allows the tools of perturbation theory to be applied to obtain convergence results for the eigenfunctions and eigenprojections. The idea is to consider the general case where \(\kappa_j \in \mathbb{C}\). Then \((\kappa_j)\) are singularities of the resolvent function: \(R : \mathbb{C} \rightarrow \mathbb{C}, R(z) := \frac{1}{\kappa - zI}\), where \(I\) is the identity operator. Write \(\hat{\kappa}_j\) as \(\hat{\kappa}_j^n\) to emphasize its dependence on \(n\). Then as \(n \rightarrow \infty\), the sequence of eigenvalue estimators \((\hat{\kappa}_j^n)_{n \in \mathbb{N}}\) converges to the singularity \(\kappa_j\). Since convergence for all \(j\) is at uniform rate, after some large enough \(n\), each sequence \((\hat{\kappa}_j^n)\) lies almost surely in a circle \(C_j^\epsilon\) of radius \(\epsilon\), and the circles for different \(j\) are disjoint.
Therefore, functional calculus methods can be applied to the resolvent function (for example, integrating around a closed contour inside $C_{\epsilon}^j$), and from there bounds can be obtained. This is the underlying idea of spectral perturbation analysis, and is used repeatedly by Dauxois et al. [12], Cardot [7] to obtain convergence and central limit theorems for estimators. We state below a result for convergence of the eigenvalues and eigenprojections, obtained from this method by Dauxois.

Lemma 17. (Dauxois et al.) [12]
Write the eigenprojection estimator $\hat{\Pi}_j^n$ and the eigenfunction estimator $\hat{\psi}_j^n$ respectively, to emphasize its dependence on $n$. For each $j \in \mathbb{N}$ and each $n \in \mathbb{N}$, $\hat{\Pi}_j^n$ is a random variable from $(\Omega, \mathcal{A}, P)$ into $(\mathcal{S}, \mathcal{B}_S)$. The sequence $(\hat{\Pi}_j^n)_{n \in \mathbb{N}}$ converges to $\Pi_j$ in $\mathcal{S}$ almost surely. If the eigenvalue $\kappa_j$ is of multiplicity 1, then $\hat{\psi}_j^n$ is a well-defined random variable from $(\Omega, \mathcal{A}, P)$ into $L^2(\mathcal{I})$. The sequence $(\hat{\psi}_j^n)_{n \in \mathbb{N}}$ converges to $\psi_j$ in $L^2(\mathcal{I})$ almost surely. The subset of $\Omega$ on which the eigenvalues, the eigenfunctions and the eigenprojections converge is the same as the subset of $\Omega$ on which $\hat{K}_n$ converges to $K$.

Another important implication of proposition 16 is: if $E\|X\|^4 < \infty$, then by lemma 14, we have mean-squared convergence at the “parametric” rate $O(n^{-1})$ for both the eigenvalue and the eigenfunction estimators, despite the infinite dimensionality of the data. Proposition 16 also illustrate the eigenspacings problem: at order $O(n^{-1/2})$, eigenspacings affects the convergence of the eigenfunction estimators $\hat{\psi}_j$, but not that of the eigenvalue estimators $\hat{\kappa}_j$. We shall see when analyze a theorem by Hall and Hosseini-Nasab [17] that at higher order, eigenspacings also affects the convergence of the eigenvalue estimators. In other words, eigenfunction estimators are more sensitive to variations in the covariance operator, making them more difficult to estimate in “worse” conditions, such as the case of sparse data, which will be analyzed in chapter 3.

1.6.3 Asymptotics of the eigenvalue and eigenfunction estimators

Given the close connection between the eigenfunctions, eigenvalues and the covariance operator, we would expect asymptotically the distribution of $\hat{\kappa}_j$ and $\hat{\psi}_j$ to be related to the distribution of $Z$ (recall, $Z$ was defined in lemma 15). Note that $Z$ itself is the kernel of an operator, say, $\mathcal{K}$, in $\mathcal{S}$. Then $\mathcal{K}$ is a Gaussian random
element in $\mathcal{S}$, with covariance function that of $(X - \mu)^* \otimes (X - \mu)$. That is,

$$
K_{\mathcal{S}} = \sum_{i,j,k,l \in \mathbb{N}} \sqrt{\kappa_i \kappa_j \kappa_k \kappa_l} E(\xi_i \xi_j \xi_k \xi_l)(\psi^*_i \otimes \psi_j) \otimes (\psi^*_k \otimes \psi_l) \\
- \sum_{i,j \in \mathbb{N}} \kappa_i \kappa_j (\psi^*_i \otimes \psi_i) \otimes (\psi^*_j \otimes \psi_j)
$$

(1.4)

This covariance function induces a covariance operator, which we shall also denote by $K_{\mathcal{S}}$ to reduce notational load. Then we have the following result.

**Proposition 18. Asymptotic distribution of the eigenvalues and eigenfunctions** (Dauxois et al.) [12]

Assume that the eigenvalues are all distinct. For each $j \in \mathbb{N}$, the asymptotic distribution of $n^{1/2}(\hat{\kappa}_j^n - \kappa_j)$ is the distribution of the eigenvalue of $\Pi_j K \Pi_j$, which is a Gaussian random variable with mean zero. Define the operator

$$
S_j := \sum_{l \neq j} (\kappa_j - \kappa_l)^{-1} \psi^*_l \otimes \psi_l.
$$

Then for each $j \in \mathbb{N}$, $n^{1/2}(\hat{\psi}_j^n - \psi_j)$ converges in distribution to the zero-mean gaussian random variable $S_j K(\psi_j)$.

The proof, which is based on spectral perturbation and convergence of measures, is omitted and can be found in Dauxois et al. [12]. To illustrate the implications of the theorem, we shall consider the case where the eigenvalues are all distinct, and $X$ is a Gaussian process. This example comes from the results in Dauxois et al. [12].

**Example 19.** (Dauxois et al.) [12]

Suppose that $X$ is a Gaussian process. Then the standardized principal components $(\xi_i)$ are not only uncorrelated, but also independent and distributed as $N(0, 1)$. This implies

$$
E(\xi_i \xi_j \xi_k \xi_l) = \begin{cases} 
3 & \text{if } i = j = k = l \\
1 & \text{if } i = j, k = l, i \neq k; \text{ or } i = k, j = l, i \neq j; \text{ or } i = l, j = k, i \neq j \\
0 & \text{otherwise.}
\end{cases}
$$
Apply this to the formula for the covariance obtained in equation (1.4). Then the covariance operator $K_K$ now has the formula

$$
K_K = 2 \sum_j \sum_{i<i,j} \kappa_i \kappa_j \phi_{ij}^* \otimes \phi_{ij} + 2 \sum_i \kappa_i^2 \phi_{ii}^* \otimes \phi_{ii},
$$

where $\phi_{ij} = \frac{1}{\sqrt{2}}(\psi_i^* \otimes \psi_j + \psi_j^* \otimes \psi_i)$ for $i < j$, and $\phi_{ii} = \psi_i^* \otimes \psi_i$. Under this notation, we recognize the family $\{\phi_{ij} : i, j \in \mathbb{N}\}$ as the orthonormal eigenbasis of the covariance operator $K_K$. Hence, apply the Karhunen-Loève expansion to the process $K$, we can write: [12]

$$
\mathcal{K} \overset{a.s.}{=} \sqrt{2} \sum_j \sum_{i<j} \sqrt{\kappa_i \kappa_j} \eta_{ij} \phi_{ij} + \sqrt{2} \sum_i \kappa_i^2 \eta_{ii} \phi_{ii},
$$

where the $(\eta_{ij})$ are independent variables distributed as $N(0, 1)$. Under this representation of $\mathcal{K}$, direct computation gives

$$
\Pi_j \mathcal{K} \Pi_j = 2^{1/2} \kappa_j \eta_{ii} \phi_{ii}, \quad S_j \mathcal{K}(\psi_j) = \sum_{k \neq j} \sqrt{\kappa_j \kappa_k} \frac{1}{\kappa_j - \kappa_k} \eta_{jk} \psi_k.
$$

Therefore, the eigenvalue estimator $\hat{\kappa}_j$ is asymptotically normal, with distribution

$$
n^{1/2}(\hat{\kappa}_j - \kappa_j) \overset{d}{\to} N(0, 2\kappa_j^2).
$$

The asymptotic distribution of $\hat{\psi}_j$ lies in the orthogonal complement of the space $\Pi_j(L^2(\mathcal{T}))$. The marginal distribution in each direction $\psi_k$ is normal, with variance $\frac{\kappa_j \kappa_k}{(\kappa_j - \kappa_k)^2}$. It is clear that the direction $\psi_j$ with the most variation is where the eigenspacings $\kappa_j - \kappa_k$ is the smallest. Again, this demonstrates the effects of eigenspacings on the convergence of $\hat{\psi}_j$. \[\Box\]

When $X$ is a Gaussian process, from equation (1.5), it is easy to construct a symmetric confidence interval for the eigenvalue $\kappa_j$. Dauxois et al. [12] showed that for $\kappa_j \neq \kappa_k$, the random elements $n^{1/2}(\hat{\kappa}_j - \kappa_j)$ and $n^{1/2}(\hat{\kappa}_k - \kappa_k)$ are asymptotically independent. This implies that if the eigenvalues are all distinct, we can construct joint confidence sets for the eigenvalues. The case for eigenfunctions is more complicated. Even when $X$ is a Gaussian process, the asymptotic distribution of $n^{1/2}(\hat{\psi}_j - \psi_j)$ contains infinitely many unknown eigenvalues. Further, Dauxois et

\footnote{Here we abuse the notation and write $\eta_{ik}$ for $\eta_{kj}$ where $k < j$.}
al. [12] showed that for $\psi_j \neq \psi_k$, $n^{1/2}(\hat{\psi}_j - \psi_j)$ and $n^{1/2}(\hat{\psi}_k - \psi_k)$ are not asymptotically independent. Therefore, constructing joint confidence sets from proposition 18 is difficult.

1.6.4 Eigenspacings and confidence intervals

For a general process $X$, it is difficult to derive an explicit formula for the process $K$, and hence confidence intervals for both eigenvalues and eigenfunctions are difficult to obtain from proposition 18. Hall and Hosseini-Nasab [17] suggested a bootstrap approach to this problem. The theoretical justification of the method relies on a Taylor-like expansion of $\hat{\psi}_j$ and $\hat{\kappa}_j$ into terms of order $O(n^{-1/2})$, $O(n^{-1})$, $O(n^{-3/2})$, ..., which is obtained under some assumptions on the smoothness of $X$ and the decay rate of the eigenvalues $\kappa_j$. Similar work can be found in the finite dimensional context, where Besse [2] applied a jacknife procedure in constructing confidence intervals for the eigenvectors, based on a Taylor-type expansion for the projections $\hat{\Pi}_j$. However, the proof methods for the two papers are different: Besse [2] relied on spectral perturbation theory, while Hall and Hosseini-Nasab [17] used moment bounds. The proof idea in [17] can also be applied to other situations, such as sparse FPCA. Intermediate results from the proof of the theorem are also used in deriving bounds for $\hat{\psi}_j$ and $\hat{\kappa}_j$ in the functional linear regression context [16].

We first state the theorem and analyze some of its implications, and conclude with a summary of the proof.

Notations and assumptions:

Let $Z := n^{1/2}(\hat{K} - K)$, $\rho_j := \min(\kappa_j - \kappa_{j+1}, \kappa_{j-1} - \kappa_j)$. Let $\xi_j := \inf \frac{\kappa_k}{\kappa_j} \left(1 - \frac{\kappa_k}{\kappa_j}\right)$, $\eta_j := \inf \frac{\kappa_k}{\kappa_j} - 1$. Let $s_j := \sup \frac{|\psi_j(u)|}{u \in I}$.

Assume the following:

- The eigenvalues ($\kappa_j$) are all distinct.
- Smoothness condition: for each $C > 0$ and some $\epsilon > 0$,
  \[ \sup_{t \in I} E\{|X(t)|^C\} < \infty, \]
  and,
  \[ \sup_{s,t \in I} E\left[\{|s-t|^{-\epsilon}|X(s) - X(t)|\}^C\right] < \infty. \]
- Moment condition: for each integer $r \geq 1$, $\kappa_{j}^{-T} E(\langle X, \psi_j \rangle)^{2r}$ is bounded uniformly in $j$. 


Call these assumptions “condition C”. The smoothness condition requires that the process $X$ has all moments finite for all $t$ and the sample paths have some Holder-type continuity. This is a strong smoothness assumption on $X$, much stronger than the existence of 4th order moment assumption, for example. Hall and Hosseini-Nasab [18] pointed out that if $X$ is left-continuous (or right-continuous) at each point with probability 1, and if that smoothness assumption hold, then for each $C > 0$

$$E(\|\hat{K} - K\|_S^C) < (\text{some constant}) \times n^{-C/2}.$$ 

By the reverse triangle inequality, this implies that $\|K\|$ is bounded in all orders. This is a quite strong assumption on the covariance operator. The moment condition ensures that all moments of the squared standardized principal components are uniformly bounded in $j$. A process that satisfies this property is the Gaussian process, for example.

**Example 20.** Condition C holds if $X$ is a Gaussian process with Holder continuous sample paths.

**Proof:** The smoothness condition is automatically satisfied. We only need to check that $\kappa_j^{-r}E(\langle X, \psi_j \rangle)^{2r}$ is bounded uniformly in $j$ for each integer $r$. Indeed, since $X$ is Gaussian, $\langle X, \psi_j \rangle$ is a normal random variable, mean zero, variance $\kappa_j$. Then,

$$\kappa_j^{-r}E(\langle X, \psi_j \rangle)^{2r} = \kappa_j^{-r}Var(\langle X, \psi_j \rangle)^r = \kappa_j^{-r} \kappa_j^r = 1.$$

This completes the proof. □

Now we can state the important theorem.

**Theorem 21.** (Hall and Hosseini-Nasab) [17]: Under condition C, for each $j$ for which

$$\|\hat{K} - K\|_S \leq \frac{1}{2} \kappa_j \min(\xi_j, \eta_j),$$

we have,

$$\hat{\kappa}_j - \kappa_j = n^{-1/2} \langle Z\psi_j, \psi_j \rangle + n^{-1} \sum_{k, k \neq j} (\kappa_j - \kappa_k)^{-1}(\langle Z\psi_j, \psi_k \rangle)^2 + O_p(n^{-3/2})$$
\hat{\psi}_j(t) - \psi_j(t)
= n^{1/2} \sum_{k: k \neq j} (\kappa_j - \kappa_k)^{-1} \psi_k(t) \langle Z \psi_j, \psi_k \rangle - \frac{1}{2} n^{-1} \psi_j(t) \sum_{k: k \neq j} (\kappa_j - \kappa_k)^{-2} \langle Z \psi_j, \psi_k \rangle^2 
+ n^{-1} \times \sum_{k: k \neq j} \psi_k(t) \left\{ (\kappa_j - \kappa_k)^{-1} \sum_{l: l \neq j} (\kappa_j - \kappa_l)^{-1} \langle Z \psi_j, \psi_l \rangle \langle Z \psi_k, \psi_l \rangle 
- (\kappa_j - \kappa_k)^{-2} \langle Z \psi_j, \psi_j \rangle \langle Z \psi_j, \psi_k \rangle \right\} + O_p(n^{-3/2}).

We note the following implications.

- The expansion shows explicitly the effect of eigenspacings on the convergence rates of \hat{\psi}_j and \hat{\kappa}_j. In particular, eigenspacings have a first-order effect on the eigenfunction estimators \hat{\psi}_j, but only a second-order effect on the eigenvalue estimators \hat{\kappa}_j. This is consistent with the observations stated earlier.
- From the theorem, we can derive asymptotic results for the estimators. Asymptotics of order \( O(n^{-1/2}) \), of course, agree with those in proposition 18. It should be noted that the expansion for \( \hat{\psi}_j - \psi_j \) is given at a chosen point \( t \in I \). In other words, we can rewrite the expression in the theorem to obtain a convergence result for \( \hat{\psi}_j - \psi_j \) in the \( L^\infty \)-norm (that is, uniform convergence on \( I \)).
- Since the eigenvalues are distinct, \( \kappa_j \min(\xi_j, \eta_j) = \rho_j \). So the condition \( \| \hat{K} - K \|_S \leq \frac{1}{2} \kappa_j \min(\xi_j, \eta_j) \) implies that expansion applies for eigenvalues \( (\kappa_j) \) with adequate spacing relative to the accuracy of \( \hat{K} \). Since \( \hat{K} \) converges to \( K \) almost surely, as \( n \) increases, the expansion can be applied for more eigenvalues. Therefore, condition \( \| \hat{K} - K \|_S \leq \frac{1}{2} \kappa_j \min(\xi_j, \eta_j) \) can be interpreted as a bound on \( \| \hat{K} - K \|_S \) relative to the eigenspacings around \( \kappa_j \), rather than a restriction on which eigenvalues and eigenfunctions the expansion can be applied for.
- Hall and Hosseini-Nasab [17] suggested bootstrapping techniques to find confidence intervals for the estimators, which use the theorem for theoretical justification. We shall not go into the details here. Discussions and simulation results can be found in [17].

Proof of the theorem:

The idea of the proof is to break up the term \( \hat{\psi}_j - \psi_j \), and bound the components using the smoothness and moment assumptions. Both the expansion for \( \hat{\kappa}_j \) and \( \hat{\psi}_j \) were derived this way. The argument, and especially the lemmas derived as stepping stones for the proof, can also be used in other contexts, for example, functional linear regression [17, 16]. Therefore, we shall give a brief outline of the proof here. There are two main steps.
Step 1: expand the term $\hat{\psi}_j - \psi_j$ in the $(\psi_k)$ basis. The result contains the term $(\hat{\kappa}_j - \kappa_k)^{-1}$. Use the geometric series on this term to obtain

$$(\hat{\kappa}_j - \kappa_k)^{-1} = \sum_{s=0}^{\infty} (\kappa_j - \hat{\kappa}_j)^2 (\kappa_j - \kappa_k)^{-(s+1)}.$$  

This allows a series expansion for $\hat{\psi}_j - \psi_j$, with terms involving spacing of eigenvalues.

$$\hat{\psi}_j - \psi_j = \kappa_j^{-1} \sum_{s=0}^{\infty} \left( \frac{\kappa_j - \hat{\kappa}_j}{\kappa_j} \right)^s \sum_{k:k \neq j} \psi_k \left( (\hat{K} - K) \hat{\psi}_j, \psi_k \right)$$

$$+ \sum_{s=0}^{\infty} (\kappa_j - \hat{\kappa}_j)^s \sum_{k:k \neq j} \left\{ (\kappa_j - \kappa_k)^{-(s+1)} - \kappa_j^{-(s+1)} \right\} \times \psi_k \left( (\hat{K} - K) \hat{\psi}_j, \psi_k \right)$$

$$+ \psi_j \left( \hat{\psi}_j - \psi_j, \psi_j \right). \quad (1.6)$$

From here, to derive the expansion for the eigenvalues, expand the term $(\hat{\kappa}_j - \kappa_j)(\hat{\psi}_j, \psi_j)$ and bound it to get

$$\left| (\hat{\kappa}_j - \kappa_j)(1 + \langle \chi_j, \psi_j \rangle) - \langle (\hat{K} - K)(\psi_j + \chi_j), \psi_j \rangle \right| \leq \left\| \hat{K} - K \right\|_S (|\hat{\kappa}_j - \kappa_j| + \left\| \hat{K} - K \right\|), \quad (1.7)$$

where $\hat{\psi}_j - \psi_j = \chi_j + \delta_j$. $\chi_j$ can be thought of as the “major contribution” term, and $\delta_j$ is the remainder. This particular step is often used as the starting point for other proofs, for example, in functional linear regression [16].

Step 2: choose the appropriate $\chi_j$ to obtain new bound for equation (1.7). The choice of $\chi_j$ clearly depends on the context and the smoothness assumptions made on $X$. In this theorem, initially $\chi_j$ is chosen to be the $1$-partial sum in the expansion (1.6). Then this term is broken down in order to replace the term $\langle (\hat{K} - K)\hat{\psi}_j, \psi_k \rangle$ in equation (1.6) by $\langle (\hat{K} - K)\psi_j, \psi_k \rangle$. This step can be thought of as quantifying the error comes from using the basis $(\hat{\psi}_j)$ instead of $(\psi_j)$. At each step of reducing to a new $\chi_j^{\text{new}}$, bound for $\chi_j^{\text{old}} - \chi_j^{\text{new}}$ is derived using the moments and smoothness assumptions. Therefore, successively tighter bounds for equation (1.7) are obtained, until we have

$$\left| \hat{\kappa}_j - \kappa_j - \langle (\hat{K} - K)\psi_j, \psi_j \rangle - \sum_{k:k \neq j} (\kappa_j - \kappa_k)^{-1} (\langle (\hat{K} - K)\psi_j, \psi_k \rangle)^2 \right| \leq D_{nj}.$$  

Where $D_{nj}$ is a random variable, $D_{nj} = O_p(n^{-3/2})$. This leads to the stochastic expansion of $\hat{\kappa}_j - \kappa_j$. Expansion in higher order terms can be obtained by taking $\chi_j$ to be the $p$-partial sum in the expansion (1.6).

The proof of $\hat{\psi}_j - \psi_j$ proceeds similarly.
Chapter 2

Functional linear model

Functional linear model (FLM), also known as functional linear regression, is a linear regression model where the response is scalar, but the predictor and the slope are functions. Such models arise in practical applications in the field of chemometrics [13], climatology [31] and neurology [11] to name a few. Since the slope “parameter” is a function, in a nonparametric context, it is determined by an infinite number of unknowns. This raises an ill-posed problem for the multivariate approach, where each measurement is treated as an observation. To take a concrete example, consider the Canadian weather dataset of Ramsay and Silverman [31].

The predictor contains $35 \times 365$ measurements, $X_{ij}$, of temperature recorded at the $i^{th}$ station on the $j^{th}$ day. The response $Y_i$ is the total annual precipitation at weather station $i$. A naive multivariate approach would be to set up the model

$$Y_i = \beta_0 + \sum_{j=1}^{365} X_{ij} \beta_j + \epsilon_i, \quad i = 1, 2, \ldots, 35.$$ 

This is a system of 35 equations with 366 unknowns, so there are infinitely many sets of solutions, all giving perfect prediction of the observed data. The problem is more than a simple overfit: Ramsay and Silverman [31] analyzed a model with monthly average temperature rather than daily variations, reducing the problem to 13 unknowns. The authors showed that the problem raises statistical questions beyond the formal difficulty of fitting an under-determined model. And clearly, taking measurements on a sparser grid might not be the most effective dimension reduction technique.

This is where the functional viewpoint demonstrates its advantages over the multivariate alternative. Due to its wide applications, functional linear modelling is one of the areas of functional data analysis that has undergone the most development since the late 90s, both theoretically and practically. Approaches to the problem include dimension reduction using FPCA [7, 17, 16], spline and basis function expansion [8, 31], nonparametric kernel methods [13, 1]. Generalizations of functional
linear model and hypothesis testing were also investigated by various researchers. A list of references can be found in the book by Ramsay and Silverman [31].

There are more appropriate multivariate techniques which deal with highly correlated data better than the naive approach mentioned above. A variety of multivariate tools were developed by the chemometrics community. Major methods consist of partial least squares (PLS), principal components regression (PCR) and ridge regression (RR). The theoretical properties of these tools were summarized in 1993 by Frank and Friedmann [14]. However, it was noted by several authors (Marx, Eilers, Hastie, Mallows, amongst others) that a functional technique which takes into account the functional nature of the data is more beneficial than the multivariate approaches mentioned above [8]. The contexts considered by these authors are often practice-based, in the sense that the function $X$ is assumed to be observed as a collection of discrete points on a grid. Therefore, the main approaches proposed often include smoothing methods, such as $B$-spline smoothing. Computational aspects of these methods were discussed in detail in Ramsay and Silverman [31], chapter 15. Some theoretical discussions can be found in Cardot et al. [8]. As mentioned, more recent developments include nonparametric approaches, and FLM is an active area of research.

In this chapter, we shall only study the theoretical properties of the functional linear model under the assumption that the whole curve $X$ is observed. Furthermore, we shall study FLM as an application of FPCA, therefore, only methods based on FPCA are discussed in details. We take an operator viewpoint, where FLM is shown to be equivalent to an ill-posed inverse problem. Treatment of such problems often involve regularization. Mathematical background on inverse problems, including regularization and definition of estimators are discussed in section 1. In section 2, we study one of the most common technique of regularization in this field, regularization by truncation, which is also a FPCA-based technique. The link between this method and the finite-dimensional technique PCR is briefly discussed. The choice of the smoothing parameter $m$ and consistency are the main points of focus. In section 3, we shall compare the performance of mean integrated squared error (MISE) of the estimator from the FPCA truncation method versus that obtained by Tikhonov regularization. We shall analyze results of Hall and Horowitz [16], which shows that under certain assumptions, both estimators attain the same optimal rate, though the Tikhonov one is more robust against eigenspacing.

In practice, the problem of estimating the “slope” is arguably not as important as the prediction problem of estimating $E(Y | X = x)$. While the former is a nonparametric problem, Hall and Horowitz [16] showed that it is possible to attain $O(n^{-1/2})$.
rate in mean-squared convergence for the predictor. Cardot [10] also noted this difference while deriving central limit theorems in the two contexts. These points shall be discussed in details in the last section.

2.1 Definitions

This introduction follows that of Cardot et al. [7, 10] and Hall and Horowitz [16], with Groetsch [15] as the main reference to the theoretical background on inverse problems.

Let \( Y : \Omega \rightarrow \mathbb{R} \) be a real-valued random variable, \( EY^2 < \infty \). Let \( X = (X(t), t \in \mathcal{I}) \) be a continuous time process defined on the same space \( (\Omega, \mathcal{A}, P) \), with finite mean trajectory \( \int_{\mathcal{I}} E(X^2) < \infty \). Let \( \mu_Y := E(Y), \mu_X := E(X) \).

**Definition 22.** A functional linear regression of \( Y \) on \( X \) is the problem of finding the solution \( b \in L^2(\mathcal{I}) \) which solves the minimization problem:

\[
\inf_{\beta \in L^2(\mathcal{I})} E\{Y - \mu_Y - \langle \beta, X - \mu_X \rangle \}^2.
\]

Some simple calculation shows that a solution \( b \) exists and is unique if and only if \( b \) satisfies

\[
Y - \mu_Y = \langle b, X - \mu_X \rangle + \epsilon,
\]

where \( \epsilon : \Omega \rightarrow \mathbb{R} \) is a real random variable with mean zero, variance \( \sigma^2 \), and is uncorrelated to \( X \), that is: \( E\{\epsilon \langle X - \mu_X \rangle \} = 0 \). Next, we introduce a measure of the relationship between \( X \) and \( Y \).

**Definition 23.** Define the cross-covariance function of \( Y \) and \( X \) to be the function \( g \in \mathcal{I} \) such that

\[
g = E\{(Y - \mu_Y)(X - \mu_X)\}.
\]

\( g \) induces a kernel operator \( G \), called the cross-covariance operator of \( Y \) and \( X \), \( G : L^2(\mathcal{I}) \rightarrow \mathbb{R} \), and

\[
G = E\{(Y - \mu_Y)(X - \mu_X)^*\}.
\]

Let \( K \) denote the covariance operator of \( X \), with eigenvalues \( (\kappa_j) \) and eigenfunctions
(ψ_j). It is easy to see that b satisfies equation (2.2) if and only if it also solves
\[ Kb = g. \] (2.3)

The problem of estimating b from K and g is known as an inverse problem in operator theory [15]. The idea is: given the observed data g and the operator K, we want to find the parameter b which gave rise to it. In this thesis, we shall work with equation (2.3) more often than equation (2.1) or (2.2) to make use of the rich theory behind inverse problems. We first introduce some mathematical background on this topic.

2.1.1 Some background on inverse problems

Let D(T), R(T), N(T) denote the domain, range and nullspace of an operator T, respectively. Let \( H^\perp \) denote the orthogonal complement of the space H. Existence and uniqueness of a solution to equation (2.3) is summarized by Picard’s condition.

**Theorem 24. Picard’s condition** [15] (p78)

Let \( g_j \) denote \( \langle g, \psi_j \rangle \), \( b_j \) denote \( \langle b, \psi_j \rangle \). Equation (2.3) has a solution if and only if \( g \in R(K) \) and for all \( \kappa_j \neq 0 \),
\[ \sum_{j=1}^{\infty} \kappa_j^{-2} g_j^2 < \infty. \] (2.4)

Further, a solution has the form
\[ b = \sum_{j=1}^{\infty} \kappa_j^{-1} g_j \psi_j + \varphi, \]

where \( \varphi \in N(K) \).

Note that the condition \( g \in R(K) \) is equivalent to \( Y \) lies in \( R\{(X - \mu_X)^*\} \oplus R\{(X - \mu_X)^*\}^\perp \), which is a dense subset of the space of measurable functions from \( \Omega \) to \( \mathbb{R} \) with finite second moment. Therefore, we can require in our model that \( g \in R(K) \). To ensure uniqueness, we need to assume that \( N(K) = \{0\} \). Note that this is equivalent to having all eigenvalues \( \kappa_j \) of \( K \) greater than 0. The theory for the case of \( N(K) \neq \{0\} \) can be dealt with by considering least-squares minimal-norm solution to equation (2.3), but in the end the same formula for the estimator is produced, so we shall not consider this case here. Further, if \( N(K) \neq \{0\} \), the least-squares minimal norm for equation (2.3) does not solve equation (2.2), which is the original problem of motivation. We summarize the discussion in this section below.
Remark 25. From now on, we assume that $g \in R(K)$, $g$ and $K$ satisfy the Picard’s condition in equation (2.4), and that $N(K) = \{0\}$. Then there exists a unique solution to equation (2.3), which is also the unique solution to equation (2.2).

Under these assumptions, we see that $K$ has an inverse $K^\dagger$ which maps $g$ to $b$. That is,

$$K^\dagger : R(K) \subseteq L^2(\mathcal{I}) \to L^2(\mathcal{I}), \quad K^\dagger(g) := \sum_{j=1}^{\infty} \kappa_j^{-1} g_j \psi_j.$$

Existence and uniqueness is not enough for an inverse problem to be called “well-posed”. For this, we need another condition: stability of the solution. This means a small perturbation in $g$ would correspond to a small perturbation in $b$, in other words, $b$ varies continuously with $g$.

Definition 26. [27] The operator equation (2.3) is said to be well-posed relative to the space $L^2(\mathcal{I})$ if for each $g$, equation (2.3) has a unique solution which depends continuously on $g$. Otherwise, it is said to be ill-posed.

The continuous dependent on $g$ implies that $K^\dagger$ is continuous. Note that $\overline{R(K)} = N(K)^\perp$ (see, for example, [15] p77). So the assumption $N(K) = \{0\}$ implies that $R(K) = D(K^\dagger)$ is dense in $L^2(\mathcal{I})$, that is, $K^\dagger$ is a densely defined operator. The following proposition gives conditions under which such an operator is bounded, or equivalently, continuous.

Proposition 27. [15] (p81)

$K^\dagger$ is a closed densely defined linear operator which is bounded if and only if $R(K)$ is closed. For $K$ a compact operator, $R(K)$ is closed if and only if $R(K)$ is finite dimensional.

Since we assumed $N(K) = \{0\}$ to obtain uniqueness, $R(K)$ cannot be finite dimensional. Therefore, proposition 27 implies that the unique solution given by $b_j = \kappa_j^{-1} g_j$ is unstable, as illustrated in the following example from Groetsch [15]:

Example 28. (Groetsch) [15] (p83)

Write

$$K^\dagger(g) = \sum_{j=1}^{\infty} \kappa_j^{-1} g_j \psi_j = \sum_{j=1}^{\infty} \kappa_j^{-1} \langle g, \psi_j \rangle \psi_j.$$
Consider an $\epsilon$-size perturbation in $g$: $g^\epsilon := g + \epsilon \kappa_j$. Then, as $j \to \infty$,

$$\|K^\dagger g - K^\dagger g^\epsilon\| = \kappa_j^{-1} \epsilon \to \infty.$$ 

This shows the idea of a small perturbation in $g$ leads to a large (in this case, infinity-size) variation in $b$.

Note that unlike the finite-dimensional case, equation (2.3) gives an ill-posed problem in the infinite dimensional context. Here the distinction lies in the fact that compact operators in a infinite-dimension space behave differently from those in finite-dimension one, as pointed out by proposition 27. This is where the theory of "functional" settings diverge from that of its multivariate parallel.

A survey of methods for ill-posed inverse problems can be found in Groetsch [15]. Here we only consider in details the method based on principal components, which is regularization by truncation. Later in this thesis, we consider Tikhonov regularization to compare the performance of estimators. The main idea of most methods for ill-posed problems is regularization: to exchange an exact, but unstable solution of the ill-posed problem (2.3) for a less exact, but stable solution of a near-by well-posed problem. This somewhat resembles the variance-bias tradeoff. The well-posed problem can be written in the general form

$$K_m b_m = g_m \tag{2.5}$$

where $m$ is a regularization parameter such that as $m$ approaches some limit value, such as infinity, $g_m \to g$, $K_m \to K$. Then for each fixed $m$, we obtain an inverse $K_m^\dagger$ such that $K_m^\dagger \to K^\dagger$. Hence $(b_m) = (K_m^\dagger g_m)$ is a sequence of stable estimators that approach $b$.

Note that stability of the solution to the well-posed problem in equation (2.5) means that $b_m$ varies continuously with perturbations in $g_m$, not with perturbations in $g$. Therefore, the choice of $m$ is a delicate matter: as $m \to \infty$, $b_m$ is a more accurate estimator for $b$, but equation (2.5) mirrors the instability of the original equation (2.3), hence $b_m$ varies more for small perturbations in $g$. The optimal choice of $m$ is based on the level of error we expect to see in $g$, and this is called a regular choice.

**Definition 29.** Regular choice of $m$ [15] (p87)

Let $g^\delta$ be an element in $L^2(\mathcal{I})$ such that $\|g - g^\delta\| \leq \delta$. Let $b_m$ and $b^\delta_m$ be the solutions to equation (2.5) with values $g_m$ and $g_m^\delta$ respectively. A choice $m = m(\delta)$
is said to lead to a regular algorithm for the ill-posed problem in equation (2.3) if

\[ m(\delta) \to \infty \quad \text{and} \quad b^\delta_{m(\delta)} \to b \quad \text{as} \quad \delta \to 0. \]

The choice of \( m \) to ensure regularity play an important role in determining convergence rate of the estimator, and this shall be our focus point of discussion. Finally, note that unlike usual inverse problems in literature (for example, those studied in Groetsch [15]), the operator \( K \) is unknown and needs to be estimated from the data. Therefore, we shall consider the general idea of estimation under regularization next.

### 2.1.2 Estimators

Assume that we observed \( n \) independent random data points \((X_i, Y_i)\), identically distributed as \((X, Y)\). Let \( \bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i, \bar{Y} := \frac{1}{n} \sum_{i=1}^{n} Y_i. \)

**Definition 30.** The estimator for the covariance function \( g \) is

\[ \hat{g} := \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) \otimes (Y_i - \bar{Y}). \]

Let \( \hat{g}_j \) denote \( \langle \hat{g}, \hat{\psi}_j \rangle \). Let \( \hat{K} \) denote the estimator for \( K \) as in chapter 1. Then, if \( b \) is the solution of equation (2.3), direct computation shows that \( b \) satisfies

\[ \hat{K}b = \hat{g} + U, \quad U := \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(\epsilon_i - \bar{\epsilon}). \]

As noted in chapter 1, the range of \( \hat{K} \) is at most of dimension \( n \). So for a fixed \( n \), no inverse of \( \hat{K} \) exists, therefore we cannot approximate \( b \) by \( \hat{b} \) defined as \( \hat{K}\hat{b} = \hat{g} \).

However, consider the empirical approximation to the regularization equation (2.5), that is, the equation

\[ \hat{K}_m\hat{b}_m = \hat{g}_m. \]

Provided that \( m \) is chosen such that \( \hat{g}_m \in R(\hat{K}_m) \), this equation is well-posed. Therefore, to find an estimator for \( b \), the idea is to construct a sequence of inverse \( \hat{K}_m \) as approximations to \( K_1 \), with \( m = m(n) \) chosen such that \( \hat{K}_m \to K_1 \) as \( n \to \infty \). Then each \( \hat{K}_m \) yields a \( \hat{b}_m \), and as \( n \to \infty \), \( \hat{b}_m \to b \). The scheme can be
represented in the following picture.

\[
\begin{align*}
K^\dagger &\quad \hat{g} \rightarrow g \quad b \\
\uparrow &\uparrow &\uparrow &\uparrow \\
\hat{K}_m^\dagger &\approx K_m^\dagger \quad \hat{g}_m \approx g_m \quad \hat{b}_m \approx b_m.
\end{align*}
\] (2.6)

Here, the arrow \(\rightarrow\) indicates that the estimators approach their true values as \(m = m(n) \to \infty\), and the \(\approx\) sign indicates that the left-hand-side quantity is an empirical approximation of the one on the right-hand-side. \(\hat{g}\) is included in the diagram since it is well-defined, and in theorems, it is often easier to work the error \(\|\hat{g} - g\|\) rather than \(\|\hat{g}_m - g\|\).

**Note:** to avoid confusion between the function \(\hat{b}_m\) and the \(m^{th}\) coefficient \(\kappa_m^{-1}\hat{g}_m\), the later shall be mentioned explicitly that it is a coefficient where used. In general, \(\hat{b}_m\) exclusively refers to the function above, unless otherwise indicated.

This is the overall idea of obtaining empirical estimators from regularization. Now we shall consider a specific method, regularization by truncation. This FPCA-related method commonly appear in the functional linear regression literature, for example, in [7, 10, 16, 31].

### 2.2 Regularization by truncation

Recall that \(K^\dagger g\), the unique solution to equation (2.3), has the formula

\[
b = K^\dagger(g) = \sum_{j=1}^{\infty} \kappa_j^{-1} g_j \psi_j.
\]

A straightforward approach to computing \(K^\dagger g\) is to truncate this decomposition at some value \(m\), which gives the estimate

\[
b_m = \sum_{j=1}^{m} \kappa_j^{-1} g_j \psi_j.
\]

To see how this is related to a well-posed inverse problem, define \(P_m\) to be the projection onto the subspace of \(L^2(I)\) spanned by the first \(m\) eigenfunctions of \(K\). That is:

\[
P_m = \sum_{j=1}^{m} \psi_j^* \otimes \psi_j.
\]

Then \(K_m := P_m K P_m : P_m(L^2(I)) \to P_m(L^2(I))\), is an operator between two finite-dimensional spaces and of full rank, hence has a continuous inverse \(K_m^\dagger\) by
proposition 27. Therefore, the problem of finding $b_m$ such that
\[ K_m b_m = P_m g =: g_m \]
is a well-posed inverse problem, with solution
\[ b_m = (K_m)^{-1} g_m = K_m^* g_m = (P_m K P_m)^{-1} P_m g, \]
as noted.

### 2.2.1 Choice of regularization parameter

Consider the problem of a regular choice of $m$. Suppose we have a perturbed version $g^\delta$ of $g$, such that $\|g - g^\delta\| < \delta$ for some fixed $\delta$. By triangle inequality,
\[ \|b_m^\delta - b\| \leq \|b_m^\delta - b_m\| + \|b_m - b\|. \]
By orthogonality of the eigenfunctions, we see that $\|b_m - b\| \leq O(\kappa_m)$, and
\[ \|b_m - b_m^\delta\|^2 = \sum_{j=1}^m (g - g^\delta, \psi_j)^2 \kappa_j^{-2} \leq \kappa_m^{-2} \sum_{j=1}^m (g - g^\delta, \psi_j)^2 \leq \delta^2 \kappa_m^{-2} \]
where the last inequality follows by the Cauchy-Schwarz inequality. So we have two important bounds
\[ \|b_m - b\| \leq O(\kappa_m), \]
and
\[ \|b_m^\delta - b\| \leq \|b_m - b\| + \delta \kappa_m^{-1} \leq \kappa_m^{-1} + \delta \kappa_m^{-1} \quad (2.7) \]
The first bound implies that a faster convergence rate of $\kappa_m$ to zero gives better accuracy for $b_m$ as an approximation of $b$. The second bound implies that for $m = m(\delta)$ to give a regular algorithm as defined in 29, $m$ needs to be chosen such that $\kappa_m^{-1} \delta \to 0$ as $\delta \to 0$, in other words, $\delta$ serves as an upper bound for the rate at which $\kappa_m$ can decay to ensure regularity. So the accuracy-stability trade-off in $b_m$ lies in the rate at which $\kappa_m(\delta)$ converges to 0 relative to $\delta$, which is the size of the perturbation in $g$. This is a very important remark, and we shall see its role in the next section.

### 2.2.2 Estimators and relation to PCR and FPCA

Now we consider the problem of estimating $K_m$ and $g_m$ from the data. Note that the empirical estimators have been defined in section 2.1.2. First, we rewrite the model
in equation (2.2) to an equivalent formulation which is more familiar to statisticians

\[ Y = a + \langle b, X \rangle + \epsilon \]

where \( a := \mu_Y - \langle b, \mu_X \rangle \) is a constant in \( \mathbb{R} \). This presentation shows that \( a \) can be interpreted as the "intercept", and \( b \) the "regression slope".

Let \( p \) denote the maximum number of eigenvalues \( \hat{\kappa}_j \) of \( \hat{K} \) that is non-zero (so \( p \leq n \)). Then for \( m \leq p \), we have the estimators:

\[
\hat{P}_m = \sum_{j=1}^{m} \hat{\psi}^*_j \otimes \hat{\psi}_j, \quad \hat{K}_m = \hat{P}_m \hat{K} \hat{P}_m, \quad \hat{g}_m = \sum_{j=1}^{m} \hat{g}_j \hat{\psi}_j.
\]

Therefore, the estimator \( b_m \) is:

\[
\hat{b}_m = \sum_{j=1}^{m} \hat{\kappa}^{-1}_j \hat{g}_j \hat{\psi}_j
\]

And the formula for \( a \) suggests the estimator

\[
\hat{a} = Y - \langle \hat{b}_m, X \rangle.
\]

In practice, the function \( \hat{b}_m \) is often identified with its vector of coefficients \( \hat{b}_m = [\hat{b}_1, \ldots, \hat{b}_m] \), and the pair \((\hat{a}, \hat{b}_m)\) is often derived as a least-squares estimator [17].

**Lemma 31.** [17]

For a fixed \( m \leq p \), let \( \hat{b}_m \) denote the vector of coefficients \([\hat{b}_1, \ldots, \hat{b}_m]\). Then the \((m + 1)\)-dimensional vector \((\hat{a}, \hat{b}_m)\) defined as above satisfies

\[
[\hat{a}, \hat{b}_m] = \text{argmin}_{[\alpha, \beta]} \sum_{i=1}^{n} \left( Y_i - \alpha - \sum_{j=1}^{m} \beta_j \eta_{ij} \right)^2,
\]

where \( \eta_{ij} = \langle X_i, \hat{\psi}_j \rangle \), \( \beta = (\beta_1, \ldots, \beta_m) \) is a \( m \)-dimensional vector.

**Outline of proof**

A detailed proof for the lemma can be found in Hall and Hosseini-Nasab [17], or the first edition of Ramsay and Silverman [31], chapter 10. The idea is to note that the above problem can be written entirely in matrix form. Let \( \eta \) denote the \( n \times m \) matrix with entries \( \eta_{ij} \), let \( Y = (Y_1, \ldots, Y_n)^T \). Let \( \overline{\eta}_j = \frac{1}{n} \sum_{i=1}^{n} \eta_{ij} \), and let \( \overline{Y} \) denote the \( n \times 1 \) matrix with \( ij^{th} \) entry \( \overline{\eta}_j \), \( Y \) denote the \( n \times 1 \) vector with \( ij^{th} \) entry \( Y_i \).
Then the least-squares solution satisfy:

$$\alpha = \bar{Y} - \sum_{j=1}^{m} b_j \bar{\eta}_j$$

and

$$(\beta_1, \ldots, \beta_m)^T = \{(\eta - \bar{\eta})^T(\eta - \bar{\eta})\}^{-1}(\eta - \bar{\eta})^T(Y - \bar{Y}).$$

Direct computation of the matrices complete the proof. See, for example, [17].

**Relation to FPCA**

Here we want to draw attention to the motivation behind choosing the eigenfunctions $\psi_j$ of $K$ as the basis for expansion in this regularization method. The idea is best illustrated via analyzing the proof of the lemma above. In this proof, note that because the basis is $\hat{\psi}_j$, the matrix $(\eta - \bar{\eta})^T(\eta - \bar{\eta})$ is diagonal, with entries $\hat{\kappa}_j$. This property greatly simplifies the formula of the least-squares solution. ($X_i$) could be expand in other basis, but then the matrix $(\eta - \bar{\eta})^T(\eta - \bar{\eta})$ would not be diagonal, hence the formula for the least-squares solution would be complicated. Indeed, the same argument justify the choice of $(\psi_j)$ as the basis for the decomposition of the operator $K$ and the function $g$ in the truncation method.

**Relation to PCR**

Principal component regression (PCR) is a multivariate method which works with a discretized version of the functional linear regression problem. By “discretized version” we mean the case where $X$ is not observed as a curve, but rather a collection of measurements on a grid. Clearly observations from one curve display high correlation, therefore a naive ordinary least squares approach would yield a bad estimate, as discussed in the introduction of this chapter. The idea of PCR is to work with the set of $n$ vectors of dimension $m \times 1$, each consist of unstandardized principal components $[\eta_1, \ldots, \eta_m]$ of the vector $X_i$. Since the principal components are uncorrelated with each other, we can do an ordinary least-squares regression of $Y_i$ on the $(\eta_{ij})$, that is, we seek the coefficients $[\alpha, \hat{b}_m]$ which minimize:

$$[\hat{a}, \hat{b}_m] = \arg\min_{[\alpha, \beta]} \sum_{i=1}^{n} \left(Y_i - \alpha - \sum_{j=1}^{m} \beta_j \eta_{ij}\right)^2,$$

The regularization parameter $m$ in PCR is seen as a smoothing parameter, and is often chosen by cross-validation [14]. Though PCR and FLM minimize the same equation, certainly the interpretation of the parameters in these two methods are different. The truncation method in FLM has more theoretical background. Note that in practice, the FLM estimator $\hat{b}_m$ is obtained by performing PCR on the
data first, then smooth the vector of coefficients using some interpolation method, such as B-spline [8]. Therefore, we can view the idea behind PCR as an alternative interpretation for lemma 31, that is, to regress \(Y\) on the first \(m\) principal components of \(X\).

### 2.2.3 Choice of \(m\) and consistency

In this section, we study the effect of \(m\) on consistency properties of \(\hat{b}_m\). We shall see the crucial role that the accuracy-stability trade-off equation (2.7) plays in determining the choice of \(m\) relative to \(n\). Under certain conditions, convergence in probability and almost surely of \(\hat{b}_m\) to \(b\) was established by Cardot et al. [7]. This paper gives important bounds and results, though the English version contains minimal discussions. Asymptotics and convergence in mean integrated squared error are postponed to later sections, since they have interesting connections with other problems.

As the diagram 2.6 suggested, we can view \(\hat{g}\) as a perturbation of \(g\), with distance \(\delta = \|\hat{g} - g\|\) from the true function \(g\). First, we need to obtain some bound on this perturbation size in terms of \(n\). Following chapter 1, we want to quantify the effect of estimating \(X\) and \(Y\) separately. Therefore, define

\[
\tilde{g} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y).
\]

Then

\[
\|g - \hat{g}\|^2 = \|g - \tilde{g}\|^2 + \|\tilde{g} - \hat{g}\|^2 + 2\|g - \tilde{g}\|\|\tilde{g} - \hat{g}\|.
\]

Using the same proof idea as lemma 12 in chapter 1, we can show that \(\|\tilde{g} - \hat{g}\| = O(n^{-1})\). The following lemma gives a bound for \(\|g - \tilde{g}\|\):

**Lemma 32.** (Cardot et al.) [7]

If \(E\|X\|^4 < \infty\), then:

\[
E\|g - \tilde{g}\|^2 \leq \frac{\|b\|^2 E\|X - \mu_X\|^4}{n} + \frac{\sigma^2}{n} E\|X - \mu_X\|^2.
\]

Therefore,

\[
E\|g - \tilde{g}\|^2 \leq \frac{\|b\|^2 E\|X - \mu_X\|^4}{n} + \frac{\sigma^2}{n} E\|X - \mu_X\|^2 + O(n^{-2}).
\]

**Proof:** By definition of the \(L^2\)-norm in \(L^2(\mathcal{I})\), for any orthonormal basis \((e_j)\) of

---

39
In $L^2(I)$, we have:
\[
\|\hat{g}\|^2 = \sum_{j=1}^{\infty} \langle g, e_j \rangle^2.
\]
Expand $E\|\hat{g}\|^2$, and use the independence of the $(X_i)$, we obtain
\[
E\|\hat{g}\|^2 = \sum_{j=1}^{\infty} \frac{1}{n} E\{\langle X - \mu_X, e_j \rangle (Y - \mu_Y)\}^2 - \frac{1}{n} \|g\|^2 + \|g\|^2.
\]
Write $Y = \mu_Y + \langle b, X - \mu_X \rangle + \epsilon$, and use Cauchy-Schwarz inequality to get:
\[
\frac{1}{n} \sum_{j=1}^{\infty} E\{\langle X - \mu_X, e_j \rangle (Y - \mu_Y)\}^2 \leq \frac{1}{n} \left( \|b\|^2E\|X - \mu_X\|^4 + \sigma^2E\|X - \mu_X\|^2 \right).
\]
And the bound for $\hat{g}$ follows. Coupled with the discussion before the lemma, we obtain the bound for $\hat{g}$.

Now write $\hat{g}$ as $\hat{g}_n$ to emphasize its dependence on $n$. We can view $(\hat{g}_n)$ as a sequence of perturbations of $g$, with distance from the true function $g$ being
\[
\delta_n := \|\hat{g}_n - g\| = O(n^{-1/2}).
\]
To bound $\|\hat{b}_m - b_m\|$, we cannot simply apply bound in equation (2.7), since we have another source of perturbation comes from estimating $K$ by $\hat{K}$. One way to bound this error is to write out the term $(\hat{P}_m \hat{K} \hat{P}_m)^{-1}$, then bounding the eigenfunction estimators using proposition 16. This is done in [7], and the authors obtained the following proposition

**Proposition 33.** (Cardot et al.) [7]
\[
\|b_m - \hat{b}_m\| \leq \gamma_n \|K - \hat{K}\|_{\infty} + \frac{1}{\hat{\kappa}_m} \|g - \hat{g}_n\| + O(n^{-1})
\]
where $\gamma_n = 2\|g\| \left\{ \frac{1}{\hat{\kappa}_m n_m} + 2\left( \frac{1}{n_m} + \frac{1}{\hat{\kappa}_m} \right) \sum_{j=1}^{m} a_j \right\}$, with $a_j := \frac{2\sqrt{2}}{\rho_j}$, where $\rho_j$ was defined in proposition 16.

The version stated above has an extra term $O(n^{-1})$ compared to the version in [7], which comes from estimating the mean. This does not affect the results at order $n^{-1/2}$, and in particular, the theorems on convergence almost surely and in probability in [7] hold without modifications.
So the accuracy-stability trade-off equation for the sequence of estimators \((\hat{b}_m)\) is
\[
\|\hat{b}_m - b\| \leq O(\kappa_{m+1}) + \gamma_n\|K - \hat{K}\|_{\infty} + \frac{1}{\kappa_m}\|g - \hat{g}\| + O(n^{-1}).
\]

For \(m = m(n)\) to be a regularization parameter on an event \(E_n\), the right-hand side needs to approach 0 as \(n \to \infty\). Suppose, for a moment, that on an event \(E_n\),
\[
E(\|K - \hat{K}\|_{\infty}|E_n) = O(E\|K - \hat{K}\|_{\infty}), \quad E(\|g - \hat{g}_n\||E_n) = O(E\|g - \hat{g}\|),\]
and that \(C_1\hat{\kappa}_m \leq \hat{\kappa}_m \leq C_2\hat{\kappa}_m\) for some constants \(C_1, C_2 < \infty\). Then using the bounds in lemma 14 and proposition 32, \(E(\|\hat{b}_m - b\||E_n)\) can be bounded by
\[
O(\kappa_{m+1}^2) + \frac{O(n^{-1})}{\kappa_m^4} + \frac{O(n^{-1})}{\kappa_m^2 \left(\sum_{j=1}^{m} a_j\right)^2} + \frac{O(n^{-1})}{\kappa_m^2}.
\]

This bound would converge to 0 if, for example, we had condition \(C_1\):

\[
\begin{cases}
\lim_{n \to \infty} n\kappa_m^2 = \infty \\
\lim_{n \to \infty} \frac{n\kappa_m^2}{(\sum_{j=1}^{m} a_j)^2} = \infty.
\end{cases}
\]

Cardot et al. [7] considered the event
\[
E_n = \{\frac{\hat{\kappa}_m}{2} < \hat{\kappa}_m < \frac{3\hat{\kappa}_m}{2}\}.
\]

On \(E_n\), \(\gamma_n \leq v_n = \|g\| \left(\frac{4}{\kappa_m^4} + \frac{6}{\kappa_m \sum_{j=1}^{m} a_j}\right)\). Working with the tail probability rather than the expectation directly, we obtain
\[
P(\|b - \hat{b}_m\| > \eta, E_n) \leq P\left(\|K - \hat{K}\|_{\infty} > \frac{\eta}{2v_n}\right) + P\left(\|g - \hat{g}_n\| > \frac{\kappa_m\eta}{4}\right).
\]

To pass from tail probabilities to expectation, apply Chebychev’s inequality for the right-hand side. Then we obtain an upper bound for \(P(\|b - \hat{b}_m\| > \eta, E_n)\) in terms of \(E\|K - \hat{K}\|^2\) and \(E\|g - \hat{g}_n\|^2\). As discussed above, applying the bounds in lemma 14 and 32, we see that condition \(C_1\) is sufficient for this probability to converge to 0. A bound for \(P(\overline{E_n})\) comes from lemma 14, and hence under condition \(C_1\), this quantity also converges to 0. This is the essence of the proof of the following result, which was first stated in [7].

**Theorem 34.** (Cardot et al.) [7]
Let \(a_j\) be defined as above. Let \((m(n))_{n \in \mathbb{N}}\) be a sequence of positive integers that approaches \(\infty\) as \(n \to \infty\), with \(m(n) \leq n\). Assume the following:
• $\kappa_1 > \kappa_2 > \ldots > 0$
• $\hat{\kappa}_1 > \hat{\kappa}_2 > \ldots > \hat{\kappa}_m > 0$
• $E\|X\|^4 < \infty$.

• Condition $C_1$:  
  \[
  \begin{cases} 
  \lim_{n \to \infty} n\kappa_m^4 = \infty \\
  \lim_{n \to \infty} \frac{n\kappa_m^2}{(\sum_{j=1}^{m} a_j)^2} = \infty.
  \end{cases}
\]

Then:  
\[
\|\hat{b}_m - b\| \to 0 \quad \text{in probability as } n \to \infty.
\]

Comments: We note that the assumptions of the theorem are really basic: they ensure that the estimator $\hat{b}_m$ is well-defined and regularized. The first condition eliminate any pathologies arising from $a_j$ blowing up because of ties in the eigenvalues sequence, the second condition ensures that each $\hat{b}_j$ is well-defined, the third condition allows the use of lemma 14 and 32, and the last condition ensures regularity as discussed.

The bounds in lemma 14 and 32 are rather crude. Indeed, if we impose some moment assumptions, almost sure convergence can be obtained for a slightly faster convergence rate of the eigenvalues, while maintaining the regularization property.

Theorem 35. (Cardot) [7] Assume the following:

• $\kappa_1 > \kappa_2 > \ldots > 0$
• $\hat{\kappa}_1 > \hat{\kappa}_2 > \ldots > \hat{\kappa}_m > 0$
• $\|X\| \leq c_1 \text{ a.s. (That is, as a random variable mapping to } L^2(I), X \text{ is bounded almost everywhere.)}$
• $|\epsilon| \leq c_2 \text{ a.s. (This means that the error is bounded almost everywhere)}$

• Condition $C_2$:  
  \[
  \begin{cases} 
  \lim_{n \to \infty} \frac{n\kappa_m^4}{\log(n)} = \infty \\
  \lim_{n \to \infty} \frac{n\kappa_m^2}{(\sum_{j=1}^{m} a_j)^2 \log(n)} = \infty.
  \end{cases}
\]

Then:  
\[
\|\hat{b}_m - b\| \to 0 \quad \text{almost surely as } n \to \infty
\]

The proof of the above theorem can be found in [7]. We note that if the eigenvalues $\kappa_j$ decay geometrically or exponentially, an application of the theorems yield the following corollary.
Corollary 36. If \( m(n) = o(\log n) \), and \( \kappa_j = ar^j \) for some \( a > 0 \), \( r \in (0, 1) \) or \( \kappa_j = aj^{-\gamma} \) for some \( a > 0 \), \( \gamma > 1 \), then conditions \((C_1)\) and \((C_2)\) are satisfied.

Note the slow convergence rate of \( o(\log n) \) on the sequence \( m \) for this case. If the covariance function \( K \) is smooth (that is, the eigenvalues \( \kappa_j \) decay fast to 0), then the regularity condition requires that \( m \) converges rather slowly. In practical terms, a slowly increasing \( m \) implies that we need a very large increase in the sample size \( n \) before we can add another term to the approximation \( \hat{b}_m \). Intuitively this leads to a poor convergence rate for \( \hat{b}_m \). This rate can be improved if we work with a smooth \( b \), and if \( b \) is sufficient smooth relative to \( K \), some optimal convergence rate can be achieved. This is the idea behind the theorem by Hall and Horowitz [16] which gives optimal mean integrated squared error converge rate of \( \hat{b}_m \) within a certain class of estimators.

2.3 Mean integrated squared error

An important measure of the “goodness” of an estimator is the mean integrated squared error (MISE) [23], which is defined as

\[
MISE(\hat{b}_m) = E \int_I (\hat{b}_m - b)^2 = E \|\hat{b}_m - b\|^2.
\]

We recognize that convergence in MISE is convergence in mean-square. Statistically, a result on bounds or convergence MISE of \( \hat{b}_m \) is of great interest.

Under some smoothness and eigenvalue-decay assumptions, Hall and Hosseini-Nasab [17] obtained:

\[
\int_I E(\hat{b}_m - b)^2 \sim \frac{\sigma^2}{n} \sum_{j=1}^{m} \kappa_j^{-1} + \sum_{j=m+1}^{\infty} b_j^2,
\]

where \( \sigma^2 = Var(\epsilon) \), and \( A_n \sim B_n \) means that the ratio of the random variables \( A_n \) and \( B_n \) converges to 1 as \( n \to \infty \). The eigenvalues \( \kappa_j \) here are assumed to be decaying at a polynomial rate which favors a faster convergence of \( \hat{b}_m \). The second term on the right-hand side is the bias of the estimator \( \hat{b}_m \), and the first term is interpreted by the authors as the dominant term in the contribution of the error \( \hat{b}_m - b \) to the integrated squared error. This term clearly shows the effect of the decay rate of \( \kappa_j \), but the lack of some type of \( O(n^\alpha) \) bound overall makes the result difficult to interpret.
Hall and Horowitz [16] worked with exponentially decaying eigenvalues, but the authors imposed a smoothness requirement on the function \( b \) to ensure that it is sufficiently smooth relative to \( K \). Their theorem states that under the assumptions of the model, for some appropriate choices of \( m \), \( \hat{b}_m \) achieves the optimal convergence rate in MISE, which is a nonparametric rate. The authors also considered an alternative regularization method, Tikhonov regularization, which is not based on FPCA. It was shown that the same optimal rate is achieved under similar assumptions. This implies that slow convergence rate is an intrinsic property of the problem, rather than a property of the estimator.

We first introduce Tikhonov regularization for the purpose of comparing the optimal MISE rate of its estimator with that obtained by truncation. Since this is the only time we shall mention Tikhonov regularization in the thesis, only a brief summary will be provided. Tikhonov regularization is an important method in solving inverse problems. An introduction can be found in chapter 5 of Groetsch [15].

### 2.3.1 Tikhonov regularization

Here our notations and definitions follow that of Groetsch [15], chapter 5. Definitions of the estimators follow that of Hall and Horowitz [16].

Consider the original ill-posed inverse problem

\[
Kb = g.
\]

In Tikhonov regularization, we want to approximate this with the equation

\[
(K + \lambda I)b_\lambda = g, \tag{2.8}
\]

where \( \lambda > 0 \), \( I \) is the identity operator. Since \( K \) is a bounded positive semidefinite operator, the operator \( (K + \lambda I) \) is bounded and strictly positive, therefore it has a bounded inverse. So provided that \( g \in R(K + \lambda I) \), the inverse problem in equation (2.8) is well-defined, and we have the solution

\[
b_\lambda = (K + \lambda I)^{-1}g.
\]

It can be shown that \( b_\lambda \to b \) as \( \lambda \to 0 \). Here \( \lambda \) is the regularization parameter (that is, it plays the role of \( m \)). If \( \delta \) denote a perturbation in \( g \), then

\[
\|b_\lambda^\delta - b\| \leq O(\lambda) + \lambda^{-1/2}\delta.
\]

\(^1\)which we can assume it to be, because \( N(K + \lambda I) = \{0\} \) hence \( R(K + \lambda I) \) is dense in \( L^2(\mathcal{I}) \)
This is the accuracy-stability trade-off equation for Tikhonov regularization, the analogue of equation (2.7) in the truncation case.

In the FLM context, estimators in Tikhonov regularization are defined as follows: the estimator for \((K + \lambda I)^\dagger\) is

\[
\hat{K}^+ := (\hat{K} + \lambda I)^{-1}.
\]

This makes the estimator for \(b_\lambda\) to be

\[
\tilde{b}_\lambda = \hat{K}^+ \hat{g}
\]

where \(\hat{g}\) and \(\hat{K}\) are defined previously (in section 1).

It is important to note that the operator \(K + \lambda I\) is not compact, therefore we do not have the same kind of spectral representation as for \(K\). In particular, the estimating of \(\tilde{b}_\lambda\) does not involve any basis expansion. Instead, the inverse \((K + \lambda I)^\dagger\) is computed directly from the data, so we obtain a sequence of inverses \((K^\dagger_\lambda) := ((K + \lambda I)^\dagger)\) that converges to \(K^\dagger\) as \(\lambda \to 0\). It is this property of Tikhonov regularization that makes the estimator robust to the eigenspacing problem.

### 2.3.2 MISE of the truncation method

Assume the following:

- Moment assumptions on \(X\): \(X\) is assumed to have finite fourth-order moment, in the sense that \(E\|X\|^4 < \infty, E\langle X, \psi_j \rangle^4 < C\kappa_j^2\) for all \(j\), (where \(C > 1\) denote a constant). The errors \(\epsilon_i\) are independent and identically distributed, zero mean, \(\sigma^2 = Var(\epsilon_i) \leq C\).
- Decay rate of \(\kappa_j\) and \(b_j\): for some \(\alpha > 1, \beta > 1 + \frac{\alpha}{2}\):

\[
\kappa_j - \kappa_{j+1} \geq C^{-1} j^{-\alpha-1} \text{ for } j \geq 1, \quad |b_j| \leq C j^{-\beta}.
\]

Call these assumptions \((A)\). Let \(\mathcal{F}(C, \alpha, \beta)\) denote the set of distributions of \((X, Y)\) that satisfy assumption \((A)\), for given values of \(C, \alpha, \beta\). Let \(\mathcal{B}\) denote the class of measurable functions \(\tilde{b}\) of the data \((X_1, Y_1), \ldots, (X_n, Y_n)\), generated by equation (2.2). Then the theorem states that if we choose the truncation parameter \(m\) correctly, that is, if \(m\) satisfies:\(^2\)

\[
m \asymp n^{1/(\alpha+2\beta)},
\]

\(^2\)where the relation \(r_n \asymp s_n\) for positive \(r_n\) and \(s_n\) means that the ratio between \(r_n\) and \(s_n\) is bounded away from zero and infinity.
then the resulting estimator \( \hat{b}_m \) is optimal in the class \( \mathcal{B} \), in the sense that it achieves the fastest rate of convergence for the MISE in this context.

**Theorem 37.** (Hall and Horowitz) [16]
Assume the conditions (A) and that \( m \gg n^{1/(\alpha+2\beta)} \). Then

\[
\lim_{D \to \infty} \limsup_{n \to \infty} \sup_{F \in \mathcal{F}} \left\{ \int I(\hat{b}_m - b)^2 > D n^{-2(\beta-1)/(\alpha+2\beta)} \right\} = 0 \tag{2.9}
\]

as \( n \to \infty \). Furthermore, for any \( \tilde{b} \in \mathcal{B} \):

\[
\liminf_{n \to \infty} n^{(2\beta-1)/(\alpha+2\beta)} \inf_{\tilde{b} \in \mathcal{B}} \sup_{F \in \mathcal{F}} \int E_F(\tilde{b} - b)^2 > 0.
\]

The authors of the theorem noted that with some truncation of the term \( \hat{b}_m \) to prevent it from taking too large values, the first equation of the theorem implies that:

\[
\sup_{F \in \mathcal{F}} \int E_F(\tilde{b}_m - b)^2 = O(n^{-(2\beta-1)/(\alpha+2\beta)}), \tag{2.10}
\]

where \( \tilde{b}_m \) denote the truncated version of \( \hat{b}_m \). This can be defined as follows [5]: for some arbitrary chosen, fixed positive constant \( C_3, C_4, C_5 \), let

\[
\tilde{b}_m = \begin{cases} 
\hat{b}_m, & \text{if } \|\hat{b}_m\| \leq C_4 n^{C_5} \\
C_3, & \text{otherwise}
\end{cases}
\tag{2.11}
\]

Equation (2.10) shows that the MISE of \( \tilde{b}_m \) converges to 0 at rate \( O(n^{-(2\beta-1)/(\alpha+2\beta)}) \). The second equation in the theorem shows that any estimator \( \bar{b} \) in \( \mathcal{B} \) can only obtain the rate \( O(n^{-(2\beta-1)/(\alpha+2\beta)}) \) at best. Therefore \( \hat{b}_m \) attains the optimal MISE convergence rate in the class \( \mathcal{B} \).

**Outline of proof**

This is a rather technical proof, therefore we shall consider the main points only. The idea is to quantify the errors from approximating \( \psi_j, \kappa_j \) and \( g_j \) separately, give bounds on these errors, and add up together to get a convergence rate for the MISE. The authors defined \( \tilde{b}_j = \tilde{\kappa}_j^{-1} g_j \), so \( \hat{b}_j - \tilde{b}_j \) is the error from estimating \( g_j \) by \( \hat{g}_j \). This error can be broken into three components: \( S_{j1} = \langle \hat{g} - g, \psi_j \rangle \), which is the error from estimating \( g \) by \( \hat{g} \), \( S_{j2} = \langle g, \hat{\psi}_j - \psi_j \rangle \), which is the error from estimating \( \psi_j \) by \( \hat{\psi}_j \), and \( S_{j3} \) is the cross-term \( S_{j3} = \langle \hat{g} - g, \hat{\psi}_j - \psi_j \rangle \). Then under the assumptions of the theorem, the authors obtained a bound for this term. Now, \( \tilde{b}_j - b_j \) is the error...
comes from estimating $\kappa_j$ by $\hat{\kappa}_j$. Under assumptions of the theorem, and using some bounds obtained from theorem 21, the authors showed that

$$\sum_{j=1}^{m} (\hat{b}_j - b_j)^2 = o_p(n^{-(2\beta-1)/(\alpha+2\beta)}).$$

We have one more error, which is the bias term $\sum_{j=m+1}^\infty b_j^2$ that comes from estimating $b$ by $\hat{b}_m$. In [16], this is bounded by $O(n^{-(2\beta-1)/(\alpha+2\beta)})$. Therefore, the $O_p(n^{-(2\beta-1)/(\alpha+2\beta)})$ bound comes mainly from bounding the first term, $\sum_{j=1}^{m} (\hat{b}_j - b_j)^2$. So in some sense, the estimation of $\kappa_j$ by $\hat{\kappa}_j$ does not influence the convergence rate of $\hat{b}_m$ as much as the estimation of $g_j$ by $\hat{g}_j$. Intuitively this should be the case, as the later involves estimation of the eigenfunctions $\psi_j$.

**Discussions**

The moment assumption is not strict: a finite fourth-order moment is a basic requirement to ensure consistency of the eigenvalues, eigenfunctions and the $\hat{b}_m$. The good convergence rate is obtained not by smoothness of $K$, but by the relative smoothness between $b_j$ and $K$. The condition $\beta > 1 + \frac{1}{2} \alpha$ can be interpreted as $b$ being no less smooth than the lower bound of the smoothness of $K$, which is measured by the decay rate of its eigenvalues. As analyzed previously, this smoothness assumption of $b$ relative to $K$ allows a good convergence rate for the estimator $\hat{b}_m$, while keeping regularity.

It should be noted that the decay rate of $\kappa_j$ is not written in the form $\kappa_j \geq C j^{-\alpha}$ for a constant $C > 0$, but rather in terms of eigenspacing. Indeed, this is not equivalent: the spacing assumption implies $\kappa_j \geq C j^{-\alpha}$, but the reverse can fail, for example, when there is a sequence of ties, or closely spaced eigenvalues. A specific example can be found in [16]. This spacing requirement comes from the sensitivity of the eigenfunction estimator $\hat{\psi}_j$ to eigenspacing, as analyzed in chapter 1. We shall see in Tikhonov regularization that this eigenspacing assumption is not required, because the method does not use $\psi_j$.

The tuning parameter $m$ here is chosen to minimize the MISE, not merely to ensure regularity. Of course some regularity is implicitly implied, as a larger-than-necessary $m$ leads to near-unstable behaviour of the equation $\hat{K}_m \hat{b}_m = \hat{g}_m$, which results in a large MISE. Note that with this $m$, under the assumptions of the model, condition $C_2$ in theorem 35 is satisfied. Indeed, $m(n)$ increases much faster than the (albeit conservative) rate $o(\log n)$ suggested in corollary 36. Again, this emphasizes the importance of the relative smoothness assumption between $b$ and $K$. 

47
The optimal convergence rate under these conditions satisfies:

\[
\frac{2\beta - 1}{\alpha + 2\beta} < \frac{2\beta - 1}{4\beta - 1} < \frac{1}{2},
\]

so estimating \(b\) is a nonparametric problem. The authors of the theorem noted that this is a “typical” rate for ill-posed inverse problems [16]. This is demonstrated in the following section, where we shall see that under similar conditions, Tikhonov regularization also attains the same optimal rate.

## 2.3.3 MISE in Tikhonov regularization

In a similar context, Hall and Horowitz [16] in the same paper derived optimal MISE convergence rate for the Tikhonov-regularized estimator \(\tilde{b}\). The moment assumption is the same, however, the decay rate is replaced by the following conditions:

For some \(\alpha > 1, \beta > \alpha - \frac{1}{2}\):

\[
C\kappa_j \geq j^{-\alpha} \quad \text{for} \quad j \geq 1, \quad |b_j| \leq Cj^{-\beta}.
\]

Denote this and the moment condition by assumption \((B)\). Let \(G(C, \alpha, \beta)\) denote the set of distributions \(F\) of \((X, Y)\) that satisfy assumption \((B)\). Then we have the following theorem.

**Theorem 38.** *(Hall and Horowitz) [16]*

If assumption \((B)\) holds, and if the tuning parameter \(\lambda\) satisfies

\[
\lambda \asymp n^{-\alpha/(\alpha + 2\beta)},
\]

then

\[
\sup_{F \in G} \int_{I} E_F(\tilde{b} - b)^2 = O(n^{-(2\beta - 1)/(\alpha + 2\beta)})
\]

as \(n \to \infty\). Furthermore,

\[
\liminf_{n \to \infty} n^{(2\beta - 1)/(\alpha + 2\beta)} \inf_{b \in \mathcal{B}} \sup_{F \in G} \int_{I} E_F(\bar{b} - b)^2 > 0.
\]

The two bounds have similar interpretations to those in theorem 37. The first equation shows the performance of \(\tilde{b}\), and the second equation shows the performance of any estimator \(\bar{b}\) in the class \(\mathcal{B}\). Together, the two equations imply that \(\tilde{b}\) obtain the optimal MISE convergence rate, which is \(O(n^{-(2\beta - 1)/(\alpha + 2\beta)})\).
Discussions
The slight difference in the relation between $\beta$ and $\alpha$ between the two theorems is unimportant. If $\beta > \max(\alpha - \frac{1}{2}, \frac{1}{2}\alpha + 1)$, then the two theorems give the same optimal MISE convergence rates. Note that the decay condition in Tikhonov regularization is of the form $C\kappa_j \geq j^{-\alpha}$ rather than in terms of eigenspacing, which implies that the method is robust against closely spaced or equal eigenvalues. This is consequence of not having to estimate the eigenfunctions $\psi_j$ in Tikhonov regularization.

2.4 Selected topics in FLM

2.4.1 Prediction

In practice, estimation of $b$ is arguably not as important as the estimation of $Y$ given $X = x$. This is termed the “prediction problem”. Specifically, given that $X$ and $Y$ satisfies the functional linear model, we want to find

$$E(Y|X = x) = a + \langle b, x \rangle. \quad (2.13)$$

The theoretical aspects of the prediction problem were discussed in Cai and Hall [5], where a mean-squared convergence rate was derived. This paper considered the case of estimating $b$ by truncation, and quantified the effect of the relative smoothness between $x$, $b$ and $K$ on the convergence rate. Cardot et al. [10] derived a central limit theorem for the estimator $\hat{b}$ as well as the predictor $\langle \hat{b}, x \rangle$ for a range of regularized estimators for $b$, including those obtain by truncation and Tikhonov regularization. This paper underlined the difference between the two problems: it was shown that no central limit theorem exists for $\hat{b}$, while some asymptotic normality holds for $\langle \hat{b}, x \rangle$. Other work on the prediction problem include that of Cuevas et al. [11], where the case of $Y \in L^2(\mathcal{I})$ was considered, Ferraty and Vieu [13], where nonparametric approaches to the functional prediction problem were discussed in details. Here we shall only analyze two papers of Cai and Hall [5] and Cardot et al. [10], since they utilize the truncation estimator.

Consider the method of regularization by truncation. An estimate for equation (2.13) would be

$$\hat{a} + \langle \hat{b}_m, x \rangle = \hat{a} + \int_{\mathcal{I}} \hat{b}_m x = \hat{a} + \sum_{j=1}^{m} \hat{b}_j \bar{x}_j,$$

where $\bar{x}_j := \langle x, \hat{\psi}_j \rangle$, and $\hat{a}$ is

$$\hat{a} = \bar{Y} - \langle \hat{b}_m, x \rangle.$$
As noted in [5], the operation of integration gives additional smoothness, therefore an optimal choice of \( \langle \hat{b}_m, x \rangle \) to minimize the mean-squared error would involve an undersmoothed \( \hat{b}_m \) (relative to the optimal choice of \( \hat{b}_m \) in theorem 37). This implies a possibly better convergence rate for \( \langle \hat{b}_m, x \rangle \), and hence sets the prediction problem apart from the problem of estimating \( b \). Indeed, this distinction can be seen by comparing theorem [16] with the theorem in Cai and Hall [5]. We shall state the later theorem below.

Assume the following:

- Eigenvalues decay rate and spacing: \( C^{-1} j^{-\alpha} \leq \kappa_j \leq C j^{-\alpha} \), and \( \kappa_j - \kappa_{j+1} \geq C^{-1} j^{-\alpha-1} \) for \( j \geq 1 \).
- Moment assumption: all the moments of \( X \) are finite. Specifically, for each \( r \geq 2 \) and each \( j \geq 1 \), \( E \| \xi_j \|^{2r} < C(r) \kappa_j^r \), where \( \xi_j \) is the unstandardized principal component of \( X \). \( C(r) \) does not depend on \( j \), and for any sequence \( j_1, \ldots, j_4 \), \( E(\xi_{j_1} \xi_{j_2} \xi_{j_3} \xi_{j_4}) = 0 \) unless each index \( j_k \) is repeated. All moments of the distribution of the errors \( \epsilon_i \) are finite.
- Smoothness of \( b \): \( |b_j| \leq C_1 j^{-\beta} \) for each \( j \geq 1 \).
- Smoothness of \( x \): \( |x_j| \leq C_2 j^{-\gamma} \) for each \( j \geq 1 \).
- Relative smoothness between \( x, b \) and \( K \): \( \alpha > 1, \beta \geq \alpha + 2, \) and \( \gamma > \frac{1}{2} \).

Collectively, call the conditions above assumption (A). Define \( B(C_1, \beta) \) to be the set of \( b = \sum_{j=1}^{\infty} b_j \psi_j \) such that \( b_j \) satisfies the smoothness condition above for chosen \( C_1, \beta \). Like in theorem 37, let \( \tilde{b}_m \) denote the truncated version of \( \hat{b}_m \) (recall equation (2.11)). The convergence rate of \( \langle \tilde{b}_m, x \rangle \) to \( \langle b, x \rangle \) depends on the relation between \( \alpha \) and \( \gamma \).

**Theorem 39.** (Cai and Hall) [5]

Under assumption (A), suppose that the truncation parameter \( m \) is chosen as follows: \( m \asymp m_0 \), where

\[
m_0 = \begin{cases} 
  n^{1/2(\beta+\gamma-1)}, & \text{if } \alpha + 1 < 2\gamma, \\
  (n/\log n)^{1/(\alpha+2\beta-1)}, & \text{if } \alpha + 1 = 2\gamma, \\
  n^{1/(\alpha+2\beta-1)}, & \text{if } \alpha + 1 > 2\gamma.
\end{cases}
\]

Then, for each given \( C, C_1, \ldots, C_5 > 0 \), as \( n \to \infty \), the estimator \( \tilde{b}_m \) defined in equation (2.11) satisfies

\[
\sup_{b \in B(C_1, \beta)} E \left( \langle \tilde{b}_m, x \rangle - \langle b, x \rangle \right)^2 = O(\tau)
\]

50
where \( \tau = \tau(n) \) is given by

\[
\tau = \begin{cases} 
  n^{-1}, & \text{if } \alpha + 1 < 2\gamma, \\
  n^{-1} \log n, & \text{if } \alpha + 1 = 2\gamma, \\
  n^{-2(\beta+\gamma-1)/(\alpha+2\beta-1)}, & \text{if } \alpha + 1 > 2\gamma.
\end{cases}
\]

**Discussions**

The intercept is not taken into account here, it is noted in [5] that up to order \( O(n^{-1/2}) \) this does not change the convergence results. The second theorem in [5] shows that if \( X \) is assumed to be Gaussian, the errors \( \epsilon_i \) in the functional linear model are normal with zero mean and strictly positive variance, then the convergence rates above is optimal for \( b \in B \). We shall not analyze this theorem here, but note its rather strict assumptions to establish optimality. Comparing the assumptions of theorem 39 to those in theorem 37, we see that both the moment assumptions on \( X \) and the relative smoothness between \( b \) and \( K \) are stronger. This reflects that in worse cases (where \( b \) and \( X \) are not as smooth, for example), the optimal convergence rate in the prediction problem might not be much better from the optimal rate for \( \hat{b}_m \) in theorem [16]. Cardot [10] pointed out that the assumption \( \alpha + 1 < 2\gamma \) may be excessive. We state this as a remark below.

**Remark 40.** Let \( x \) be a realization of a random variable \( X_{n+1} \) distributed as \( X \), independent of the \((X_1, \ldots, X_n)\). By assumptions of theorem 39, since \( C^{-1}j^{-\alpha} < \kappa_j \) and \( x_j \leq C_2 j^{-\gamma} \), we have

\[
\sum_{j=1}^{\infty} \kappa_j^{-1} x_j^2 \leq \text{some constant} \times \sum_{j=1}^{\infty} j^{-(2\gamma-\alpha)}.
\]

The right-hand side series converges for \( 2\gamma - \alpha > 1 \), which implies \( \alpha + 1 < 2\gamma \). Note that \( K^{-1} \) exists, positive and self-adjoint, since we assumed that \( N(K) = \{0\} \). Therefore it has a uniquely defined square-root, denote by \( K^{-1/2} \). Now, the left-hand side can be written as

\[
\sum_{j=1}^{\infty} \kappa_j^{-1} x_j^2 = \|K^{-1/2} x\|^2.
\]

So the assumption \( \alpha + 1 < 2\gamma \) ensures that \( \|K^{-1/2} x\| < \infty \). However, \( P(\|K^{-1/2} X_{n+1}\|^2 < \infty) = 0 \) [10], so realizations \( x \) of \( X_{n+1} \) such that \( 2\gamma > \alpha + 1 \) exists with probability zero. Therefore, this smoothing assumption is indeed excessive.
Therefore, the case $2\gamma > \alpha + 1$ can be seen as being analyzed for theoretical completeness. In the two later cases where $x$ is not so smooth compared to $K$, that is, where $\alpha + 1 \geq 2\gamma$, it is clear that the choice of $m$ in this theorem approaches infinity faster than that in theorem 37, where $m$ was chosen to optimize the MISE of $\hat{b}_m$. This is consistent with the idea of under-smoothing $\hat{b}_m$ to obtain optimal convergence for the predictor $\langle \hat{b}_m, x \rangle$. Cai and Hall noted in their paper that for the case $\alpha + 1 < 2\gamma$, the choice of $m$ is less critical in obtaining the $O(n^{-1})$ convergence rate, and that the choice $m_0 = n^{1/2(\beta+\gamma-1)}$ does not reflect all the possible choices of $m$. In other words, this rate can be obtained for a wide range of $m$, so the predictor is “semirobust” against the choice of $m$. Intuitively, a smoother $x$ implies that the error in approximating $x$ by $\sum_{j=1}^{m} \langle x, \hat{\psi}_j \rangle \hat{\psi}_j$ is small, therefore, the convergence of the predictor to $\langle b, x \rangle$ can be achieved for less-smooth $\hat{b}_m$. This explains the decreasing speed of $m_0$ as $x$ becomes less smooth relative to $K$ (which is reflected via the relationship between $\alpha$ and $\gamma$).

### 2.4.2 Central limit theorems

As noted in the previous section, estimators for $b$ can have a slower convergence rate compared to estimators for $E(Y|X = x) = a + \langle b, x \rangle$. In particular, the later can attain the $O(n^{-1/2})$ convergence rate if $b$ and $x$ are sufficiently smooth, but the former is a nonparametric problem. In this section, we shall analyze a result from Cardot [10], which states that for a wide range of regularization methods, no central limit theorem holds for the estimator $\hat{b}$ of $b$, while some asymptotic normality holds for the predictor $\langle \hat{b}, x \rangle$.

The class of regularized estimators considered in [10] is quite general, covering both the truncation and Tikhonov regularized estimators. Instead of stating the exact definition, we denote a regularized estimator of this class for $b$ by $\hat{b}$. (For example, in the truncation method, $\hat{b} = \hat{b}_m$, and in Tikhonov regularization, $\hat{b} = \hat{b}_\lambda$). Then the theorem in [10] is as follow.

**Theorem 41.** (Cardot) [10]

*It is impossible for $\hat{b} - b$ to converge in distribution to a non-degenerate random element in the norm topology of the Hilbert space $\mathcal{H} = L^2(\mathcal{I})$. That is, for any normalizing sequence $\alpha_n \to \infty$, $\alpha_n(\hat{b} - b)$ does not converge in distribution in the norm topology, but to a degenerate random element.*

This important theorem highlights the nonparametric nature of the problem of estimating $\hat{b}$. It also shows that in the infinite dimensional settings, multivariate
classical results are not necessarily true any more. However, in terms of the prediction problem, some asymptotic normality results apply. The context of the theorem is as follow.

- **Moment assumption**: $X$ has finite-fourth order moment, in the sense that $\sup_l E(\xi_i^4) \leq C < \infty$ for some constant $C$, where $\xi_i$ are the standardized principal components of $X$.

- **$b \in L^1(I)$. That is, $\sum_{i=1}^{\infty} |\langle b, \psi_i \rangle| < \infty$.**

- **At least for large $j$, $\lambda_j = \lambda(j)$ for a convex position function $\lambda$.**

- **$m$ is chosen such that $\hat{b} = \hat{b}_m$ is a regularized estimator.**

Collectively, call these assumption (A). These conditions are quite unrestrictive. Note that the last condition holds for both polynomial and exponential decay.

For $x \in L^2(I)$, define

$$t_{n,x} = \sqrt{\|K^{1/2}K_n^l x\|^2}.$$ 

**Theorem 42.** (Cardot) [10]

If assumption (A) holds, and if

$$\sup_{l} \frac{|\langle x, \psi_l \rangle|^2}{l} < \infty, \quad \text{and} \quad \frac{m^3(\log m)^2}{t_{n,x}\sqrt{n}} \rightarrow 0,$$

then

$$\frac{\sqrt{n}}{t_{n,x}\sigma} \left( \langle \hat{b}, x \rangle - \langle \hat{P}_m b, x \rangle \right) \xrightarrow{w} N(0,1).$$

where $\hat{P}_m$ is the projection defined in section 2.2, and $\sigma = Var(\epsilon)$.

Note that the bias term here is random. The author of the theorem stated that no improvement is possible without further assumptions. The author also noted that if smoothness conditions are imposed so that $\|K^{-1/2} x\| < \infty$, the normalization sequence is of rate $O(\sqrt{n})$. As noted in the remark above, this smoothing is excessive, and therefore is only of theoretical interests. In other cases, the normalization sequence converges at a rate slower than $O(\sqrt{n})$. 

53
Chapter 3
Practical issues: sampling and smoothing

This section briefly consider some practical issues in implementing FPCA and their impacts on the convergence rates of estimators. In practice, input data are often given in the form of discrete points rather than a whole curve as assumed in chapter 1. Therefore, we need a procedure to convert these points to a function, which usually involve smoothing. We shall introduce some of these methods in section 1. In section 2, we consider the influence of noise on the estimators. We shall see that this effect is minor. The difficulty lies more in the number of measurements taken per curve: if the data is sparse, then performance of the estimators may be seriously affected. This phenomenon is termed “sparse data”, and shall be discussed in section 3. In the last section, we consider smoothing methods which explicitly control the smoothness of the eigenfunction estimator. This section highlights the fact that a class of FPCA models can be generated just by changing the topology of the function space.

3.1 From discrete to functional data

For simplicity, assume that we have a regular grid of \(m\) points, and we observed \(X_i(u_j)\) for \(j = 1, \ldots, m, i = 1, \ldots, n\). To perform FPCA, we need to calculate, for any pair \((u, v)\), the covariance function estimator:

\[
\hat{K}(u, v) = \frac{1}{n} \sum_{i=1}^{n} X_i(u)X_i(v).
\]

Then we want to find the eigenvalues \(\hat{\kappa}_j\) and eigenfunctions \(\hat{\psi}_j\) of \(\hat{K}\), which respectively are numbers and functions that satisfy

\[
\int \hat{K}(u, v)\hat{\psi}_j(v) = \hat{\kappa}_j\hat{\psi}_j(u).
\] (3.1)
To obtain these estimators based on the measurements observed at the $m$ points $u_1, \ldots, u_m$, there are, in general, two approaches: "interpolation" and "basis expansion". A survey of both techniques can be found in Ramsay and Silverman [31], therefore, only a brief account shall be provided here.

**Method 1: interpolation from multivariate PCA**
This is the earliest approach to FPCA, used independently by Rao and Tucker, both in 1958, who applied multivariate PCA without modification to observed function values [31]. Under the multivariate PCA viewpoint, the covariance matrix is an $m \times m$ matrix, with $\tilde{K}(u_j, u_k) = \frac{1}{n} \sum_{i=1}^{n} X_i(u_j)X_i(u_k)$.

Standard multivariate techniques yield the eigenvectors $\tilde{\psi}_j$, and the eigenvalues $\tilde{\kappa}_j$ of $\tilde{K}$. Then the function $\hat{\psi}_j$ is obtained from $\tilde{\psi}_j$ via interpolation (more accurately, by applying a standard smoothing technique such as kernel smoothing). The final formula contains a weight correction term which comes from the choice of numerical integration technique to evaluate (3.1). Details can be found in Ramsay and Silverman [31].

Clearly, when using this method, we are essentially smoothing the principle factors obtained from multivariate PCA. With this approach, in practical terms, the multivariate and the functional viewpoints do not differ significantly. The functional approach has the advantage of providing a viewpoint which has substantial theoretical background, therefore allow further developments. Indeed, as noted in chapter 2, though the estimator obtained from the truncation approach to FLM can be viewed as the result of interpolating the vector obtained from PCR. However, the theoretical background of FLM pointed out that the problem is intrinsically nonparametric, while this can be difficult to infer from the PCR viewpoint.

**Method 2: basis function expansion**
Suppose that we choose some basis $(\phi_l)$ of $L^2(\mathcal{I})$ to approximate $X$ by the $p$-term sum:

$$X \approx \tilde{X} = \sum_{l=1}^{p} \langle X, \phi_l \rangle \phi_l = \sum_{l=1}^{p} c_l \phi_l = C\Phi.$$  

Clearly $\tilde{X}$ lies in a $p$-dimensional subspace of $L^2(\mathcal{I})$, hence its covariance function $K$ can only have at most $p$ non-zero eigenfunctions. Let $\Psi$ denote the $p \times 1$ matrix
of eigenfunctions of $K$. Then $K$ and $\Psi$ can be written in matrix terms

$$K(u, v) = E\{\Phi^T(u) C^T C \Phi(v) \}, \quad \Psi = B \Phi, \quad \psi_j = b_j \Phi = \Phi^T b_j^T,$$

where $b_j$ denote the $j^{th}$ row of $B$, $B$ defined by $\Psi = B \Phi$. By identifying $\Psi$ with $B$, $X$ with $C$, we see that estimating the eigenfunctions is a $p$-dimensional problem of finding the $p \times p$ matrix $B$. Equation (3.1) now reads

$$\int K(u, v) \psi_j(v) = E\{\int \Phi^T(u) C^T C \Phi(v) \Phi^T(v) b_j^T \ dv\} = E\{\Phi^T(u) C^T C W b_j^T \} = \kappa_j \Phi^T(u) b_j^T,$$

where $W = \int \Phi(v) \Phi^T(v) \ dv$. This implies the purely matrix equation:

$$E(C^T C W b_j^T) = \kappa_j b_j^T.$$

Since $\Phi$ is known, $W$ is known, and $C$ can be approximated by numerical integration, we can solve the above equation to obtain $B$, and hence find the eigenfunctions $\Psi$. The basis expansion method is convenient for computation, since it reduces the problem directly to a matrix formula. Clearly the choice of basis ($\phi_l$) is the crucial step here. In practice, ($\phi_l$) is chosen based on data features [31].

Both methods extend naturally to the case of FLM. As mentioned in chapter 2, the interpolation method in the FLM context corresponds to performing PCR on the discretized data. Implicit smoothing is done in both approaches. In the interpolation method, numerical integration of equation (3.1) and the interpolation (or rather, smoothing) step gives a smooth estimate for $\psi_j$. In the basis expansion method, numerical integration of $C$, and the choice of $p$ correspond to a smoothing of the input data $X$, and hence result in smooth estimates of the eigenfunctions ($\psi_j$). Accuracy of the approximation depends on these steps. If the measurements are observed on a dense grid and if $X$ is smooth, a situation which usually arises when the data are recorded by some automated instrument, then the numerical approximation performs quite well, and the choice of numerical method usually does not have a great effect. However, if the data are sparse, estimators obtained this way can be highly biased. Hence a different treatment is required. We shall discuss this in section 3.
3.2 A model for noise

Suppose we are given a set $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ of independent random functions, identically distributed as $Y$, with:

$$Y(t) = X(t) + \epsilon(t),$$

where $X$ is a stochastic process defined on a compact set $\mathcal{I}$, with finite mean-trajectory: $\int_{\mathcal{I}} E(X^2) < \infty$, and the set of random variables $\{\epsilon(t) : t \in \mathcal{I}\}$ are independent, identically distributed as $\epsilon$, with $E\epsilon = 0$, $Var\epsilon = \sigma^2 < \infty$. Suppose that $X$ and $\epsilon$ are independent. $\epsilon$ represents the noise, $X$ represents the true underlying process. We are interested in estimating the true eigenfunctions and eigenvalues of $X$ from the observations ($Y_i$).

In the basis expansion method, a maximum likelihood approach is often employed to estimate the matrix $B$. Therefore the noise is often assumed to be Gaussian, and it does not significantly affect the estimates [29]. In the interpolation method, note that $E[Y(u)Y(v)] = E[X(u)X(v)] + \sigma^2 \delta_{uv}$ where $\delta_{uv}$ is the Kronecker’s delta function. So the error really only have an effect on the trace of the covariance function. In practice, for example, see [36], this is dealt with by ignoring the trace of the covariance function $K_Y$ in calculations. Then, the covariance matrix would be a $(m - 1) \times (m - 1)$ matrix, with entries:

$$\tilde{K}(u_j, u_k) = \frac{1}{n} \sum_{i=1}^{n} X_i(u_j)X_i(u_k), \quad j \neq k$$

This results in eigenfunction estimators $\tilde{\psi}_j$ being vectors of $m - 1$ dimension, and from there, the interpolation method is applied as usual.

We state our conclusion as a remark below.

**Remark 43.** In both interpolation and basis expansion methods, identically distributed, independent noise with constant variance does not make a significant difference in calculating the eigenfunctions and eigenvalues.

3.3 Sparsity

Accuracy of the numerical methods mentioned in section 1 depends on the resolving power of data, that is, the number of measurements made per curve, $m$, relative to the amount of curvature in the data [31]. For example, if the underlying curve $X$ has
many high-frequency terms, then we need more measurements per curve to achieve
the same magnitude in error, as compared to the case where \( X \) is a very flat curve
(has more low-frequency terms). Here we consider the case where the underlying
curve \( X \) is smooth, but \( m \) is small. This situation is called “sparse data”, a setting
typically seen in longitudinal studies. Direct application of either methods in section
1 would lead to bad estimates due to the high bias from evaluating the integrals
numerically. However, either methods can be adapted to this setting.

Under the basis expansion framework, if \( X \) a Gaussian process, the usual approach
in literature involves finding \( B \) using maximum likelihood. James et al. [21] con-
sidered an EM algorithm method, while Peng and Paul [29] noticed that \( B \) lies in
a Stiefel manifold\(^1\), which is a Riemann surface and hence have nicely defined tan-
gents. Therefore, the authors proposed a Newton-Raphson algorithm to maximize
the likelihood. Other papers under the basis expansion framework can be found in
Peng and Paul [29] and references therein. For the interpolation approach, kernel
smoothing is often employed to obtain a smooth estimate of the observed covariance
function, then the eigenfunctions and eigenvalues are estimated as usual [36]. Hall
et al. [20] used a local linear smoother and proved optimality of this procedure
under rather weak assumptions on the process for optimal choice of bandwidths.

We shall not analyze these results in details here. It is worth summarizing thereom
1 in [20], however, since this would give us some intuition on the effects of sparsity
on convergence rates.

**Remark 44.** Theorem 1 in [20] can be summarized as follows. Assume that we
observe \( m_i \) random measurements on the \( i^{th} \) curve, \( m_i \) fixed. Under some weak
smoothness conditions on \( X \), if the smoothing bandwidth is chosen correctly, then
the resulting eigenvalue estimator \( \hat{\kappa}_j \) is still root-\( n \) consistent, but the eigenfunction
estimator converges at a much slower rate:

\[
\| \hat{\psi}_j - \psi_j \| = O_p(n^{-2/5}).
\]

It should be noted that if the \( m_i \) measurements were fixed points, then convergence
would not take place: there are infinitely many curves in \( L^2(\mathcal{I}) \) that can produce the
same observed values. Therefore, the randomness of the sampling scheme ensured
convergence as \( n \to \infty \). However, Hall et al. [20] showed that if the data collected
on a fixed or random grid is allowed to be increasingly dense, that is, \( m \to \infty \) as

---

\(^1\) Stiefel manifold: the space of real valued matrices with orthonormal columns
\( n \to \infty \) at a rate faster than \( O(n^{-1/4}) \), then the estimators produced are root-\( n \) consistent to the estimators that we would construct if the \( X_i \) were directly observed. That is,
\[
\| \tilde{\psi}_j - \hat{\psi}_j \| = O_p(n^{-1/2}),
\]
where \( \tilde{\psi}_j \) is the estimator produced under the sparse model, and \( \hat{\psi}_j \) is the estimator produced when we observe the full curves (\( X_i \)).

As mentioned, sparse data is a situation typically seen in longitudinal studies, a field where much development has been in the parametric direction. It is only in the past few years that the parallel between functional sparse data and longitudinal data were drawn [32, 20, 36, 29]. Much research in this direction is still being done.

### 3.4 Smoothing with a roughness penalty

The aim of smoothing in FPCA is to obtain smooth estimators of the eigenfunctions (\( \psi_j \)). This can be achieved by smoothing the input data \( X_i \), smoothing the covariance function \( K \), or smoothing the eigenfunctions (\( \hat{\psi}_j \)). From this viewpoint, it is clear that some smoothing is already imposed by the conversion of the discrete measurements to a functional observation, especially in the sparse data case. However, even when the whole curve is observed, smooth estimators may be desirable. Several procedures have been proposed: kernel smoothing of the covariance function [3], ridge regression [30], change of inner product [33]. In this section, we shall consider the last two methods and their relationships. In particular, we shall demonstrate the idea that by changing in the geometry of the function space, we can obtain a range of eigenfunction estimators from the same data.

#### 3.4.1 Ridge regression

One of the first papers on ridge regression in FPCA was written by Ramsay and Silverman [30]. The idea is to view FPCA as a problem of minimizing the expected mean squared error, and attach a roughness penalty to this equation to obtain smoother estimates of the eigenfunctions \( \psi_j \). Ramsay and Silverman [31] suggested to take \( \| D^2 \psi_j \|^2 \) as the roughness penalty, where \( D \) denote the differential operator. Let \( V \) denote the subspace of \( L^2(\mathcal{I}) \) where \( \phi \in V \Rightarrow D^2\phi \in L^2(\mathcal{I}) \). Then, the procedure is to find a set of orthonormal basis \( (\phi_i)_{i=1}^M \) of \( L^2(\mathcal{I}) \) that minimize

\[
E \left\| X - \sum_{i=1}^M \langle X, \phi_i \rangle \phi_i \right\|^2 + \lambda \| D^2\phi_i \|^2, \tag{3.2}
\]

59
where $\lambda > 0$ is a smoothness parameter. As $\lambda \to \infty$, we obtain very smooth estimates since more penalty are put on the norm of the second derivative of $\phi_i$. Similarly, the estimates become rougher as $\lambda \to 0$. The first edition of the book by Ramsay and Silverman [31] contains details on implementing this method in practice. Though the idea of ridge regression is familiar to statisticians, it is difficult to integrate this method into a linear operator framework. It is also not clear if this is equivalent to first smooth the data, then perform FPCA. Indeed, there is a close connection between smoothing the input data $X$ and smoothing the eigenfunctions $\hat{\psi}_j$, and this link has been established in many different settings, for example, [31, 33, 28, 3]. Therefore, we shall consider another roughness penalty method.

3.4.2 A change in norm

Silverman [33] suggested to attach a roughness penalty into the definition of the inner product itself. The author suggested to consider functions from a subset of the Sobolev space $W^{(2,2)}$:

$$V = \{ f \in W^{(2,2)}, D^4f \in L^2(I), D^2f = D^3f = 0 \text{ at the boundary of } I \},$$

and provide this space with an inner product which is a slight generalization of the usual $W^{(2,2)}$ inner product, defined by:

$$\langle \cdot, \cdot \rangle_\alpha : V \times V \to \mathbb{R}, \quad \langle f, g \rangle_\alpha = \langle f, g \rangle + \alpha \langle D^2f, D^2g \rangle,$$

where $\langle \cdot, \cdot \rangle$ denote the usual $L^2$-norm, $D$ is the differential operator, and $\alpha$ is a chosen, fixed number, $\alpha \geq 0$. Then we perform FPCA on $V$. In other words, we want to find $M$ eigenfunction estimators $(\hat{\psi}_i)_{i=1}^M$, orthonormal in $V$, such that they are the solution to

$$\arg\min_{(\phi_i) : \langle \phi_i, \phi_j \rangle_\alpha = \delta_{ij}} E \left\| X - \sum_{i=1}^M \langle X, \phi_i \rangle_\alpha \phi_i \right\|_{\alpha}^2. \quad (3.3)$$

To see how this is related to smoothness, the key step is to establish a relationship between performing FPCA with respect to the proposed inner product and performing a certain FPCA with respect to the $L^2$ inner product. Consider the space and the inner product defined. The conditions on $V$ ensure that $D^2$ is a self-adjoint linear operator. Then, since $D^4$ maps $V$ to $L^2(I)$, we can write: $\langle D^2f, D^2g \rangle = \langle D^4f, g \rangle$. So:

$$\langle f, g \rangle_\alpha = \langle Tf, g \rangle$$
where \( T = I + \alpha D^4 \) is an operator (\( I \) is the identity operator). Further, since \( T \) is a self-adjoint, strictly positive linear operator, it has a unique self-adjoint, strictly positive square-root, denoted by \( S \). (See, for example, Wouk [35]). Note that \( S^{-1} \) is also self-adjoint. Then:

\[
\langle f, g \rangle_\alpha = \langle Sf, Sg \rangle.
\]

Now, consider the effect of performing FPCA in \( \mathcal{V} \). Recall that the FPCA eigenbasis in \( L^2(\mathcal{I}) \) can be derived as the orthonormal sequence which at the \( i^{th} \) step maximizes:

\[
E \langle X, \phi_i \rangle^2,
\]

subject to \( \| \phi_i \| = 1 \), and that \( \phi_i \) is orthogonal to all other \( \phi_j \) with \( j < i \). With the new inner product, we want to maximize:

\[
E \langle X, \phi_i \rangle^2 = E(\langle S \circ S^{-1} X, \phi_i \rangle)^2 = E(\langle S^{-1} X, S\phi_i \rangle)^2,
\]

under the orthonormal constraint: \( \langle \phi_i, \phi_j \rangle_\alpha = \langle S\phi_i, S\phi_j \rangle = \delta_{ij} \). We recognize this as equivalent to performing PCA on the transformed data \( Y = S^{-1} X \), which yields eigenfunction estimators \( \hat{\psi}_i \), and from there \( \hat{\psi}_i \) can be recovered by applying \( S^{-1} \). Silverman [33] interpreted \( S^{-1} \) as a smoothing operator. So the FPCA performed in this new space can be interpreted as PCA of smoothed data.

It can be proved quite easily that this approach maximizes [33]

\[
\frac{E(\langle X, \phi_i \rangle)}{\langle \phi_i, \phi_i \rangle + \alpha \langle D^2 \phi_i, D^2 \phi_i \rangle}.
\]

In the ridge regression method, minimizing equation (3.2) under the orthonormal constraint for the \( \phi_i \) is equivalent to maximizing:

\[
\frac{E(\langle X, \phi_i \rangle) - \alpha \langle D^2 \phi_i, D^2 \phi_i \rangle}{\langle \phi_i, \phi_i \rangle}.
\]

(3.4)

We note that the change-in-norm approach is closely related to ridge regression. This method also has a clear geometrical interpretation and is easy to implement in practice [33]. Ocaña et al. [28] and Ferraty and Vieu [13] generalized this approach into a systematic way of obtaining new FPCA estimators from the same data by changing the topology of the space. This is an active research topic in FDA.
4.1 A summary of the thesis

The idea of FPCA is to retain the first $m$ terms of the Karhunen-Loève expansion as an approximation to the data. Provided that the eigenvalues decay to zero fast enough, the functional data problem is converted to an $m$-dimension multivariate problem. If we observe the whole curve $X_i$ for each $i$, estimating the eigenvalues and eigenfunctions in FPCA are shown to be semiparametric problems: the estimators attain the optimal root-n convergence rate. The key property that set functional data apart from other multivariate high dimensional problems is that there is no “curse of dimensionality” present, due to the covariance function being fixed rather than increasing with the sample size. Theoretically, closely spaced eigenvalues is a concern, as they slow down on the convergence rate of the corresponding eigenfunctions.

In practice, closely spaced eigenvalues are rare. A more important issue of concern is the sampling rate: since we often only observe points on $X_i$ rather than the whole curve, the number of sampled points affect convergence rates. In particular, the sparser the sampling points, the worse the estimators become. If the measurements are made at fixed points on the interval, clearly no convergence can be achieved. However, if the measurements are sampled randomly rather than at fixed intervals, the estimators for the eigenvalues can be shown to converge at the parametric rate $O(n^{-1/2})$, while those for the eigenfunctions converge at a slower rate $O(n^{-2/5})$ (using a smoothed estimate for the covariance function with an appropriate choice of bandwidth). The sparse data problem is one distinct feature of functional data, and it highlights one of the differences between functional and multivariate settings.

FLM is an important area of FDA where FPCA methods can be effectively applied. Unlike the finite-dimensional case, the problem of estimating the regression slope is an ill-posed inverse problem, therefore, the estimator converges at a nonparametric rate. The convergence rate is better for the prediction problem, and under very...
strict smoothness assumptions, the predictor achieves the parametric rate. No central limit theorem exists for a wide class of estimators for the regression slope. These problems do not exist in the finite-dimensional case, since the corresponding inverse problem is well-posed. This is another important difference between functional and multivariate settings which was explored in this thesis.

4.2 Further topics

Multivariate FPCA

In this thesis we only worked with the Hilbert space $L^2(I)$, where $I$ is assumed to be a compact subset of $\mathbb{R}$. It is not difficult to extend the theory of FPCA to the case where $X$ is a random function of several variables. In fact, the setting considered by Dauxois et al. [12] was more general, with $I = (I, \mu)$ abstractly defined as a measurable space with bounded measure $\mu$. Hence, consistency and asymptotic results extend readily to this case with little changes. In practice, this extension is potentially useful in analyzing spatial data, for example. Another way to obtain a “multivariate-functional” setting is to consider the case where $X$ is a vector of random functions. One can take direct sum of Hilbert spaces to reduce the problem to the “univariate-functional” situation considered. For example, Ramsay and Silverman [31] in chapter 10 considered the case of “mixed data” where $X_1$ is a random function in $L^2(I)$, and $X_2$ is a random vector. Then in this case, we can work with random functions $X$ which take values in $\mathcal{H} = L^2(I) \oplus \mathbb{R}^d$. Again, the theory changes very little when we extend to these cases, although the practical details might be slightly different [31].

New models by changing the topology

As mentioned in the last section of chapter 3, certain methods of smoothing by roughness penalty can be incorporated into a linear operator framework. Ocaña et al. [28] considered a large class of inner products and generalized this approach. The authors obtained a relationship between performing FPCA with respect to a different inner product, and performing FPCA on transformed data in the usual $L^2$ inner product. Their paper highlights the fact that a class of inner products can give rise to a class of FPCA models for the same data. As a further generalization, Ferraty and Vieu [13] worked with semi-norm, and also noted that many new models can be constructed just by changing the topological structure of the space we work with. Model selection would be of interest in this situation, and this is another active research area in functional data analysis. [13]
Hypothesis testing
The statistical inference component of FPCA is still under research. Ferraty and Vieu [13] pointed out that hypothesis testing is an important area of functional data analysis. Some work in this direction include that of Hall and Vial [19], on testing the infinite dimensionality of the data, Cardot et al. [9], on testing of no-linear-relationship between \( Y \) and \( X \) in FLM.

An alternative to FPCA: nonparametric methods
Apart from FPCA, there are other alternative tools in analyzing FDA, some of which are outlined in the book by Ramsay and Silverman [31]. An active area of research is that of nonparametric methods, where kernel methods are extended to the functional setting. Ferraty and Vieu [13] contains an overview of both practical and theoretical aspects of this approach. The book also contains a list of open research problems, some of which have been (partly) answered in the recent years. Some papers in this direction include those of Mas [26], Berlinet et al. [1], Cardot et al. [6]. This is an active research area in functional data analysis.
References


