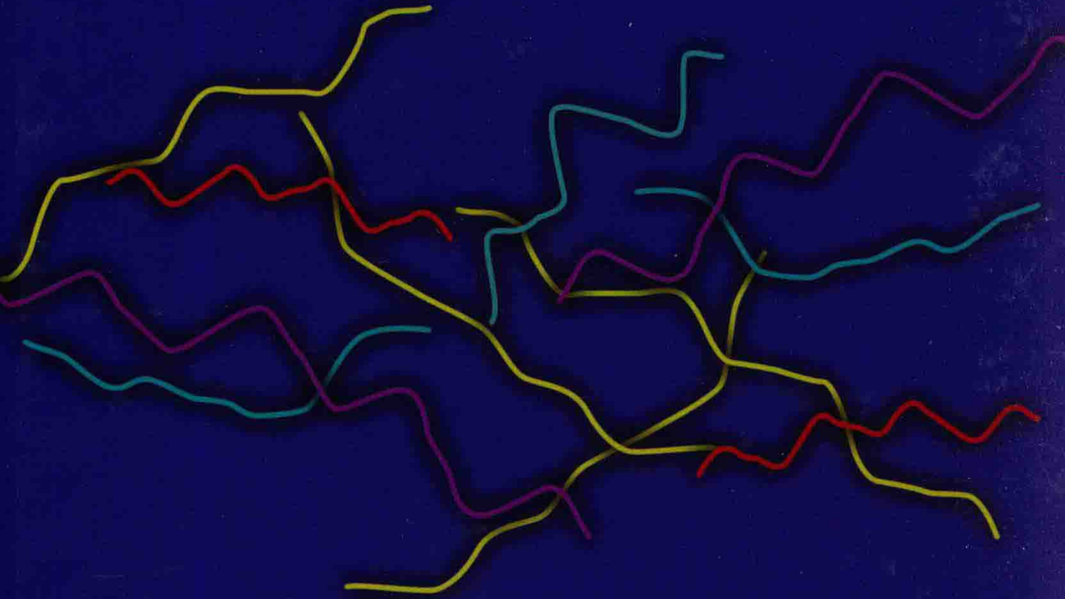
WILEY SERIES IN PROBABILITY AND STATISTICS

Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators



Tailen Hsing • Randall Eubank

WILEY

Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators

Tailen Hsing
Professor, Department of Statistics
University of Michigan, USA

Randall Eubank
Professor Emeritus, School of Mathematical and
Statistical Sciences, Arizona State University, USA

WILEY

This edition first published 2015

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data applied for.

A catalogue record for this book is available from the British Library.

ISBN: 9780470016916

Set in 10/12pt TimesLTStd by Laserwords Private Limited, Chennai, India
Printed and bound in Singapore by Markono Print Media Pte Ltd.

1 2015

Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators

WILEY SERIES IN PROBABILITY AND STATISTICS

A complete list of the titles in this series appears at the end of this volume.

To Our Families

Preface

This book aims to provide a compendium of the key mathematical concepts and results that are relevant for the theoretical development of functional data analysis (fda). As such, it is not intended to provide a general introduction to fda *per se* and, accordingly, we have not attempted to catalog the volumes of fda research work that have flowed at a brisk pace into the statistics literature over the past 15 years or so. Readers might therefore find it helpful to read the present text alongside other books on fda, such as Ramsay and Silverman (2005), which provide more thorough and practical developments of the topic.

This project grew out of our own struggle in acquiring the theoretical foundations for fda research in diverse fields of mathematics and statistics. With that in mind, the book strives to be self-contained. Rigorous proofs are provided for most of the results that we present. Nonetheless, a solid mathematics background at a graduate level is needed to be able to appreciate the content of the text. In particular, the reader is assumed to be familiar with linear algebra and real analysis and to have taken a course in measure theoretic probability. With this proviso, the material in the book would be suitable for a one-semester, special-topics class for advanced graduate students.

Functional data analysis is, from our perspective, the statistical analysis of sample path data observed from continuous time stochastic processes. Thus, we are dealing with random functions whose realizations fall into some suitable (large) collection of functions. This makes an overview of function space theory a natural starting point for our treatment of fda. Accordingly, we begin with that topic in Chapter 2. There we develop essential concepts such as Sobolev and reproducing kernel Hilbert spaces that play pivotal roles in subsequent chapters. We also lay the foundation that is needed to understand the essential mathematical properties of bounded operators on Banach and, in particular, Hilbert spaces.

Our treatment of operator theory is broken into three chapters. The first of these, Chapter 3, deals with basic concepts such as adjoint, inverse, and projection operators. Then, Chapter 4 investigates the spectral theory that underlies compact operators in some detail. Here, we present both the typical eigenvalue/eigenvector expansion for self-adjoint operators and the somewhat less common singular value expansion that applies in the non-self-adjoint case or, more generally, for operators between two different Hilbert spaces. These expansions make it possible to develop the concepts of Hilbert–Schmidt and trace class operators at a level of generality that makes them useful in subsequent aspects of the text.

The treatment of principal components analysis in Chapter 9 requires some understanding of perturbation theory for compact operators. This material is therefore developed in Chapter 5. As was the case for Chapter 4, we do this for both the self-adjoint and the-non self-adjoint scenarios. The latter instance therefore provides an introduction to the less well-documented perturbation theory for singular values and vectors that, for example, can be employed to investigate the properties of canonical correlation estimators.

The fact that sample paths must be digitized for storage entails that data smoothing of some kind often becomes necessary. Smoothing and regularization problems also arise naturally from the approximate solution of operator equations, functional regression, and various other problems that are endemic to the fda setting. Chapter 6 examines a general abstract smoothing or regularization problem that corresponds to what we call a functional linear model. An explicit form is derived for the associated estimator of the underlying functional parameter. The problems of computation and regularization parameter selection are considered for the case of real valued, scalar response data. A special case of our abstract smoothing scenario leads us back to ordinary smoothing splines and we spend some time studying their associated properties as nonparametric regression estimators.

Chapter 7 aims to establish the probabilistic underpinnings of fda. The mean element, covariance operator, and cross-covariance operators are rigorously defined here for random elements of a Hilbert space. The fda case where a random element has a representation as a continuous time stochastic process is given special treatment that, among other factors, clarifies the relationship between its covariance operator and covariance kernel. A brief foray into representation theory produces congruence relationships that prove useful in Chapter 10. The chapter then concludes with selected aspects of the large sample theory for Hilbert space valued random elements that includes both a strong law and central limit theorem.

The large sample behavior of the sample mean element and covariance operator are studied in Chapter 8. This is relevant for cases where functional data is completely observed. When meaningful discretization occurs, smoothing becomes necessary and we look into the large sample performance of two estimation schema that can be used for that purpose: namely, local linear and penalized least-squares smoothing. Chapter 9 is the principal components counterpart of Chapter 8 in that it investigates the properties of eigenvalues and eigenfunctions associated with the covariance operator estimators that were derived in that chapter.

Chapters 10 and 11 both address bivariate situations. In Chapter 10, the focus is canonical correlation and this concept is used to study a variety of fda problems including functional factor analysis and discriminant analysis. Then, Chapter 11 deals with the problem of functional regression with a scalar response and functional predictor. An asymptotically, optimal penalized least-squares estimator is investigated in this setting.

We have been fortunate to have talented coworkers and students that have generously shared their ideas and expertise with us on many occasions. A nonexhaustive list of such important contributors includes Toshiya Hoshikawa, Ana Kupresanin, Yehua Li, Heng Lian, Yolanda Munoz-Maldanado, Rosie Renaut, Hyejin Shin, and Jack Spielberg. We sincerely appreciate the invaluable help they have provided in bringing this book to fruition. The inspiration for much of the development in Chapter 10 can be traced to the serendipitous path through academics that brought us into contact with Anant Kshirsagar and Emanuel Parzen. We gratefully acknowledge the profound influence these two great scholars have had on this as well as many other aspects of our writing. TH also wishes to thank Ross Leadbetter for introducing him to the world of research and Ray Carroll for his support which has opened doors to many possibilities, including this book.

Contents

Preface **xi**

1 Introduction **1**

1.1 Multivariate analysis in a nutshell 2

1.2 The path that lies ahead 13

2 Vector and function spaces **15**

2.1 Metric spaces 16

2.2 Vector and normed spaces 20

2.3 Banach and $\mathbb{L}^p$ spaces 26

2.4 Inner Product and Hilbert spaces 31

2.5 The projection theorem and orthogonal decomposition 38

2.6 Vector integrals 40

2.7 Reproducing kernel Hilbert spaces 46

2.8 Sobolev spaces 55

3 Linear operator and functionals **61**

3.1 Operators 62

3.2 Linear functionals 66

3.3 Adjoint operator 71

3.4 Nonnegative, square-root, and projection operators 74

3.5 Operator inverses 77

3.6 Fréchet and Gâteaux derivatives 83

3.7 Generalized Gram–Schmidt decompositions 87

4 Compact operators and singular value decomposition **91**

4.1 Compact operators 92

4.2 Eigenvalues of compact operators 96

4.3 The singular value decomposition 103

4.4 Hilbert–Schmidt operators 107

4.5 Trace class operators 113

4.6 Integral operators and Mercer's Theorem 116
4.7 Operators on an RKHS 123
4.8 Simultaneous diagonalization of two nonnegative definite operators 126

5 Perturbation theory 129
5.1 Perturbation of self-adjoint compact operators 129
5.2 Perturbation of general compact operators 140

6 Smoothing and regularization 147
6.1 Functional linear model 147
6.2 Penalized least squares estimators 150
6.3 Bias and variance 157
6.4 A computational formula 158
6.5 Regularization parameter selection 161
6.6 Splines 165

7 Random elements in a Hilbert space 175
7.1 Probability measures on a Hilbert space 176
7.2 Mean and covariance of a random element of a Hilbert space 178
7.3 Mean-square continuous processes and the Karhunen–Lòeve Theorem 184
7.4 Mean-square continuous processes in $\mathbb{L}^2(E, \mathscr{B}(E), \mu)$ 190
7.5 RKHS valued processes 195
7.6 The closed span of a process 198
7.7 Large sample theory 203

8 Mean and covariance estimation 211
8.1 Sample mean and covariance operator 212
8.2 Local linear estimation 214
8.3 Penalized least-squares estimation 231

9 Principal components analysis 251
9.1 Estimation via the sample covariance operator 253
9.2 Estimation via local linear smoothing 255
9.3 Estimation via penalized least squares 261

10 Canonical correlation analysis 265
10.1 CCA for random elements of a Hilbert space 267
10.2 Estimation 274
10.3 Prediction and regression 281

10.4 Factor analysis 284
10.5 MANOVA and discriminant analysis 288
10.6 Orthogonal subspaces and partial cca 294

11 Regression 305
11.1 A functional regression model 305
11.2 Asymptotic theory 308
11.3 Minimax optimality 318
11.4 Discretely sampled data 321

References 327

Index 331

Notation Index 334

1

Introduction

Briefly stated, a *stochastic process* is an indexed collection of random variables all of which are defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$. If we denote the index set by E, then this can be described mathematically as

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

where $X(t, \cdot)$ is a $\mathscr{F}$-measurable function on the sample space Ω. The ω argument will generally be suppressed and $X(t, \omega)$ will typically be shortened to just $X(t)$.

Once the $X(t)$ have been observed for every $t \in E$, the process has been realized and the resulting collection of real numbers is called a *sample path* for the process. Functional data analysis (fda), in the sense of this text, is concerned with the development of methodology for statistical analysis of data that represent sample paths of processes for which the index set is some (closed) interval of the real line; without loss, the interval can be taken as $[0, 1]$. This translates into observations that are functions on $[0, 1]$ and data sets that consist of a collection of such random curves.

From a practical perspective, one cannot actually observe a functional data set in its entirety; at some point, digitization must occur. Thus, analysis might be predicated on data of the form

$$x_i(j/r), j = 1, \dots, r, i = 1, \dots, n,$$

involving n sample paths $x_1(\cdot), \dots, x_n(\cdot)$ for some stochastic process with each sample path only being evaluated at r points in $[0, 1]$. When viewed from this perspective, the data is inherently finite dimensional and the temptation is to treat it as one would data in a multivariate analysis (mva) context.

Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators,
First Edition. Tailen Hsing and Randall Eubank.

However, for truly functional data, there will be many more "variables" than observations; that is, $r \gg n$. This leads to drastic ill conditioning of the linear systems that are commonplace in mva which has consequences that can be quite profound. For example, Bickel and Levina (2004) showed that a naive application of multivariate discriminant analysis to functional data can result in a rule that always classifies by essentially flipping a fair coin regardless of the underlying population structure.

Rote application of mva methodology is simply not the avenue one should follow for fda. On the other hand, the basic mva techniques are still meaningful in a certain sense. Data analysis tools such as canonical correlation analysis, discriminant analysis, factor analysis, multivariate analysis of variance (MANOVA), and principal components analysis exist because they provide useful ways to summarize complex data sets as well as carry out inference about the underlying parent population. In that sense, they remain conceptually valid in the fda setting even if the specific details for extracting the relevant information from data require a bit of adjustment. With that in mind, it is useful to begin by cataloging some of the multivariate methods and their associated mathematical foundations, thereby providing a roadmap of interesting avenues for study. This is the subject of the following section.

1.1 Multivariate analysis in a nutshell

mva is a mature area of statistics with a rich history. As a result, we cannot (and will not attempt to) give an in-depth overview of mva in this text. Instead, this section contains a terse, mathematical sketch of a few of the methods that are commonly employed in mva. This will, hopefully, provide the reader with some intuition concerning the form and structure of analogs of mva techniques that are used in fda as well as an appreciation for both the similarities and the differences between the two fields of study. Introductions to the theory and practice of mva can be found in a myriad of texts including Anderson (2003), Gittins (1985), Izenman (2008), Jolliffe (2004), and Johnson and Wichern (2007).

Let us begin with the basic set up where we have a p-dimensional random vector $X = (X_1, \ldots, X_p)^T$ having (variance-)covariance matrix

$$\mathscr{K} = \mathbb{E}\left[(X - m)(X - m)^T\right] \tag{1.1}$$

with

$$m = \mathbb{E}X \tag{1.2}$$

the mean vector for X. Here, $\mathbb{E}$ corresponds to mathematical expectation and v^T indicates the transpose of a vector v. The matrix $\mathscr{K}$ admits an eigenvalue–eigenvector decomposition of the form

$$\mathscr{K} = \sum_{j=1}^{p} \lambda_j e_j e_j^T \tag{1.3}$$

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

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

where δ_{ij} is 1 or 0 depending on whether or not i and j coincide. This provides a basis for principal components analysis (pca).

We can use the eigenvectors in (1.3) to define new variables $Z_j = e_j^T(X - m)$, which are referred to as principal components. These are linear combinations of the original variables with the weight or loadings e_{ij} that is applied to X_i in the jth component indicating its importance to Z_j; more precisely,

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

In fact,

$$X = m + \sum_{j=1}^{p} Z_j e_j \tag{1.4}$$

as, if $\mathscr{K}$ is full rank, $e_1, \ldots, e_p$ provide an orthonormal basis for $\mathbb{R}^p$; this is even true when $\mathscr{K}$ has less than full rank as $e_j^T X$ is zero with probability one when $\lambda_j = 0$. The implication of (1.4) is that X can be represented as a weighted sum of the eigenvectors of $\mathscr{K}$ with the weights/coefficients being uncorrelated random variables having variances that are the eigenvalues of $\mathscr{K}$.

In practice, one typically retains only some number $q < p$ of the components and views them as providing a summary of the (covariance) relationship between the variables in X. As with any type of summarization, this results in a loss of information. The extent of this loss can be gauged by the proportion of the total X variance $V := \mathrm{trace}(\mathscr{K})$ that is recovered by the principal components that are retained. In this regard, we know that

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

while the variance of the jth component is

$$\begin{aligned} \mathrm{Var}(Z_j) &= e_j^T \mathscr{K} e_j \\ &= \lambda_j. \end{aligned}$$

Thus, the jth component accounts for $100\lambda_j/V$ percentage of the total variance and $100\left(1-\sum_{k=1}^{j}\lambda_k/V\right)$ is the percentage of variability that is not accounted for by $Z_1,\ldots,Z_j$.

Principal components possess various optimality features such as the one catalogued in Theorem 1.1.1.

Theorem 1.1.1 $\mathrm{Var}(Z_j)=\max_{\{e^Te=1,e^T\mathscr{K}e_i=0,i=1,\ldots,j-1\}}\mathrm{Var}(e^TX)$.

The proof of this result is, e.g., a consequence of developments in Section 4.2. It can be interpreted as saying that the jth principle component is the linear combination of X that accounts for the maximum amount of the remaining total variance after removing the portion that was explained by $Z_1,\ldots,Z_{j-1}$.

The discussion to this point has been concerned with only the population aspects of pca. Given a random sample $x_1,\ldots,x_n$ of observations on X, we estimate $\mathscr{K}$ by the sample covariance matrix

$$\mathscr{K}_n=(n-1)^{-1}\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)\left(x_i-\overline{x}_n\right)^T \tag{1.5}$$

with

$$\overline{x}_n=n^{-1}\sum_{i=1}^{n}x_i \tag{1.6}$$

the sample mean vector. As $\mathscr{K}_n$ is positive semidefinite, it has the eigenvalue–eigenvector representation

$$\mathscr{K}_n=\sum_{j=1}^{p}\lambda_{jn}e_{jn}e_{jn}^T, \tag{1.7}$$

where the e_{in} are orthonormal and satisfy

$$e_{in}^T\mathscr{K}_ne_{jn}=\lambda_{jn}\delta_{ij}.$$

This produces the sample principle components $z_{jn}=e_{jn}^T(x-\overline{x}_n)$ for $j=1,\ldots,p$ with $x=(x_1,\ldots,x_p)^T$ and the associated scores $e_{jn}^T(x_i-\overline{x}_n), i=1,\ldots,n$ that provide sample information concerning the Z_j.

Theorems 9.1.1 and 9.1.2 of Chapter 9 can be used to deduce the large sample behavior of the sample eigenvalue–eigenvector pairs, $(\lambda_{jn},e_{jn}),j=1,\ldots,r$. The limiting distributions of $\sqrt{n}(\lambda_{jn}-\lambda_j)$ and $\sqrt{n}(e_{jn}-e_j)$ are found to be normal which provides a foundation for hypothesis testing and interval estimation.

The next step it to assume that X consists of two subsets of variables that we indicate by writing $X=(X_1^T,X_2^T)^T$, where $X_1=(X_{11},\ldots,X_{1p})^T$ and

$X_2 = (X_{21}, \ldots, X_{2q})^T$. Questions of interest now concern the relationships that may exist between X_1 and X_2. Our focus will be on those that are manifested in their covariance structure. For this purpose, we partition the covariance matrix $\mathscr{K}$ for X from (1.1) as

$$\mathscr{K} = \begin{bmatrix} \mathscr{K}_1 & \mathscr{K}_{12} \\ \mathscr{K}_{21} & \mathscr{K}_2 \end{bmatrix}. \tag{1.8}$$

Here, $\mathscr{K}_1, \mathscr{K}_2$ are the covariance matrices for X_1, X_2, respectively, and $\mathscr{K}_{12} = \mathscr{K}_{21}^T$ is sometimes called the cross-covariance matrix.

The goal is now to summarize the (cross-)covariance properties of X_1 and X_2. Analogous to the pca approach, this will be accomplished using linear combinations of the two random vectors. Specifically, we seek vectors $a_1 \in \mathbb{R}^p$ and $a_2 \in \mathbb{R}^q$ that maximize

$$\rho^2(a_1, a_2) = \frac{\operatorname{Cov}^2(a_1^T X_1, a_2^T X_2)}{\operatorname{Var}\left(a_1^T X_1\right) \operatorname{Var}\left(a_2^T X_2\right)}. \tag{1.9}$$

This optimization problem can be readily solved with the help of the singular value decomposition: e.g., Corollary 4.3.2. Assuming that X_1, X_2 contain no redundant variables, both $\mathscr{K}_1$ and $\mathscr{K}_2$ will be positive-definite with nonsingular square roots $\mathscr{K}_i^{1/2}, i = 1, 2$. This allows us to write

$$\rho^2(a_1, a_2) = \frac{\left(\tilde{a}_1^T \mathscr{R}_{12} \tilde{a}_2\right)^2}{\tilde{a}_1^T \tilde{a}_1 \tilde{a}_2^T \tilde{a}_2}, \tag{1.10}$$

where

$$\mathscr{R}_{12} = \mathscr{K}_1^{-1/2} \mathscr{K}_{12} \mathscr{K}_2^{-1/2}, \tag{1.11}$$

$\tilde{a}_1 = \mathscr{K}_1^{1/2} a_1$ and $\tilde{a}_2 = \mathscr{K}_2^{1/2} a_2$. The matrix $\mathscr{R}_{12}$ can be viewed as a multivariate analog of the linear correlation coefficient between two variables. Using the singular value decomposition in Corollary 4.3.2, we can see that (1.10) is maximized by choosing $\tilde{a}_1, \tilde{a}_2$ to be the pair of singular vectors $\tilde{a}_{11}, \tilde{a}_{21}$ that correspond to its largest singular value ρ_1. The optimal linear combinations of X_1 and X_2 are therefore provided by the vectors $a_{11} = \mathscr{K}_1^{-1/2} \tilde{a}_{11}$ and $a_{21} = \mathscr{K}_2^{-1/2} \tilde{a}_{21}$. The corresponding random variables $U_{11} = a_{11}^T X_1$ and $U_{21} = a_{21}^T X_2$ are called the first canonical variables of the X_1 and X_2 spaces, respectively. They each have unit variance and correlation $\rho_1$that is referred to as the first canonical correlation.

The summarization process need not stop after the first canonical variables. If $\mathscr{K}_{12}$ has rank r, then there are actually $r - 1$ additional canonical variables that can be found: namely, for $j = 2, \ldots, r$, we have

$$U_{1j} = a_{1j}^T X_1 \tag{1.12}$$