Functional Data Analysis: Class notes

Alejandro Calle Saldarriaga

November 27, 2023

Contents

Abstract

This documents contain the class notes, which is more or less similar to the lectures given in person at Universidad EAFIT. The homework is taken directly from these: I propose some exercises here. I will not hand out a different file. Homework is pivotal to this course: it gives you grade and helps you prepare for exam. I recommend using the R programming language (R Core Team, 2021) (mostly because there are already many libraries out there in the CRAN which implements a lot of methods for FDA), but you are free to use whatever language you want to implement the practical exercises/homework. For Python, there is a package called scikit-fda, which is kinda recent and has many functionalities, but it is not as mature as the fda ecosystem in R. You can contact me at acalles@eafit.edu.co if you have any inquiries.

At the end of each subsection I propose some suggested readings if the material here/in class was not clear enough or you want to learn more about some specific topic.

1 Introduction

1.1 Academic Pact

- Come to class whenever possible, I believe the best way to learn in a graduate class is through discussion with your peers, and class is a great place to do that. However you can also study with this document. An ideal scenario is to attend the lectures, then read the document, then do the exercises. We will only have four in person classes, they will be online from then on. My recommended method of attending this class is the following: each week I will tell you, more or less, what I plan on teaching next week. Read up a bit on that, nothing too deep. Then, attend the lecture, ask questions based on your reading or on the class itself. Then, read carefully the topics on this document that correspond to that week of lectures before next class, and try doing the exercises. At the start of each lecture, there will be a brief discussion about exercises.
- This document will contain exercises that must be completed.
-
  - One project (50%): read two papers carefully with the same methodology (e.g. two papers on clustering for functional dataset), simulate some data and compare performances of both methods and apply both methods to a real dataset. The idea is that you will hand a document which summarized both papers and present your findings, and that you will present results to your peers in class.
  - Do the exercises in these notes (50%). This will be graded at the end of the course, however, I STRONGLY recommend doing them as you read them. Not necessarily you will be able to complete them in the first go, but you will be refining them as you learn more.
- Grading dates: To be discussed.
- **Second grader**: To request a second grader, you must follow this process.:
  1. Request the grader when the feedback for that note is done.
  2. In writing, to the teacher, in the next 3 business days following the feedback, present the request with a clear justification. The teacher has 5 business day to escalate the request accordingly.
  3. The exam will be in hands of the professor until the request is resolved.
- Bibliography: The course will mainly follow the text by Ramsay et al. (2005) and Ramsay et al. (2009) in applied aspects, and the text by Horváth and Kokoszka (2012) for a more theoretical look. See (Cuevas, 2014; Wang et al., 2016) for fairly recent overviews of the area.
- Please send typos/errors for this document acalles@eafit.edu.co.
- I'll upload the scripts I use throughout this document or in class to Interactiva.

2 A personal account of Functional Data Analysis

Functional data analysis (FDA) is an area of statistics of important current development, where the data are functions. The FDA concept was coined by Ramsay (1982), but some methods are older and date back to the 50s (Grenander, 1950; Rao, 1958). With the advance of technology, continuously recorded data have become more common, and thus interest in FDA has spiked, particularly in times of big data and data science (Aykroyd et al., 2019). In the FDA context, we consider that certain functions originated the data that we record discretely, and that those functions are the sample members, not the explicit discrete data.

So what does a typical FDA dataset looks like? A classic example is the Berkeley Growth data, which records height data for some individuals across time. See Figure 1 to visualize the dataset. These data has some interesting features: we have two groups (males vs females) we can compare, the data is observed in irregular intervals in time (observations are more frequent at earlier ages), and we have sufficiently smooth data so we can explore things like derivatives of the data.

In FDA we can do many of the same stuff we can do in statistics: We can estimate means and covariances (and do inferences about those estimated objects) (Yao et al., 2005a), regression models (Yao et al., 2005b), clustering (Li and Chiou, 2011), classification (James and Hastie, 2001), time series (Panaretos and Tavakoli,

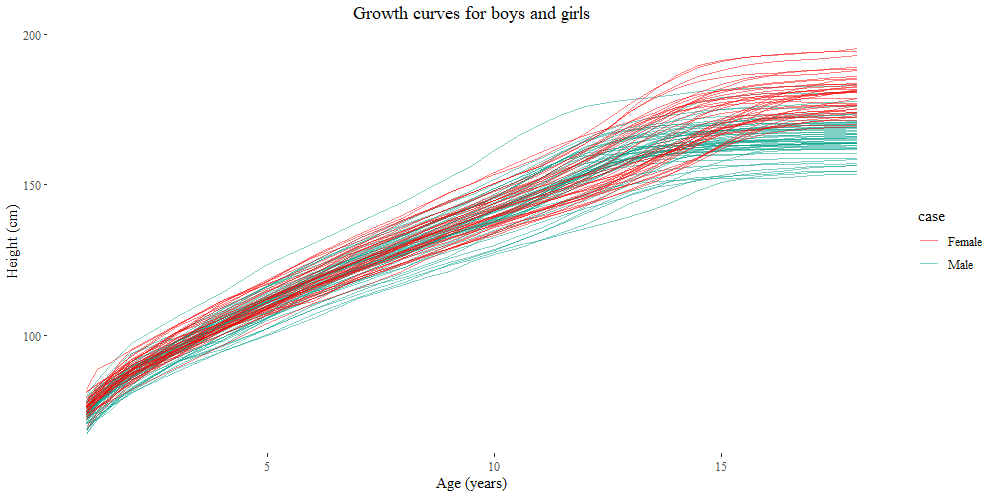


Figure 1: Berkeley growth data

2013), etc. If you can think of an area or general methodology in statistics, probably that is also generalized for functional data and you can find a few articles on it. If you can't find any articles, congratulations! You have an area of research to work on.

2.1 Motivation: Applications

FDA has been applied in many areas, and it is applied to more and more varied areas. Here I will summarize some of the findings of Ullah and Finch (2013), where they do a systematic review of applications of FDA in varied disciplines. Note that the paper is from 2013 (and only includes papers up to 2010), that means it is getting kind of old in this hugely saturated academic world. They found 84 applications of FDA, mostly from 2005 until 2010, roughly relating to the following areas:

- Biomedicine (18 papers): For example, Wu and Müller (2010) treats DNA microarray data (which is high dimensional, and in many cases, $n << p$) as a functional data object, i.e. they embed the data in functions instead of treating it as multivariate. This is followed by a functional logistic regression on those functions in order to classify gene expressions from DNA microarrays.

- Biomechanics (11 papers): For example, Durá Gil et al. (2010) compares functional and nonfunctional regression for measuring the effect of walking velocity on force platforms. They conclude that functional regression is less restrictive and more informative for this kind of data.

- Medicine (10 papers): For example, Erbas et al. (2010) uses FDA techniques to model and predict mortality rates by age in breast cancer patients.

- Linguistics (6 papers): For example, Lucero (1999) uses FDA techniques to study the acoustic components of voice signals, such as harmonics to noise ratio.

- Biology (4 papers): For example, Müller et al. (2009) studies temporal fertility trajectories of flies, and uses FDA to conclude that flies from the same species and different geographic locations have different temporal trajectories.

- Ecology (4 papers): For example, Henderson (2006) uses functional principal component analysis, functional linear models and functional cluster to monitor water quality across different sites.

- Psychology (4 papers): For example. Chapados and Levitin (2008) uses FDA to investigate physiological interactions in the listening of music. Interesting bit: James O. Ramsay, considered the father of FDA (with Bernard Silverman), was trained in Psychology. He did not receive much formal training

on math/statistics in his undergraduate years, thinking it was uninteresting and trivial, preferring to pursue areas such as literature and philosophy instead. While in Grad School for psychology he became greatly interested in statistics and the rest is history.

- Meteorology (4 papers): For example, Lee et al. (2009) uses FDA techniques to investigate how web surfing data might explain how well did certain locations respond to the risk of Huricane Katrina.

- Environmental studies (4 papers): For example, Torres et al. (2011) uses an outlier detection technique for gas emission data on the city of Oviedo, Spain. Detecting outliers here is interesting because abnormal gas emissions might be of concern for public health.

- Demography (3 papers): For example, Hyndman and Shang (2010) uses generalization of boxplots to functional data analysis applied to mortality rates. This paper is really cool! Very nice visualization tool, very useful, and very theoretically sound. The application to mortality rates almost feel like an afterthought, this tool works for mostly any functional dataset. It has its own easily usable R package, which computes truly beautiful graphs. They use French mortality graphs to detect outlying mortality curves for a given year. The outlying years were $1914 - 1919$, 1940 and $1943 - 1944$. What do those years correspond to?

  Exercise 1. *This paper is uses French male mortality curves from 1899 to 2005. An interesting exercise is to extend this analysis, at least, until 2020 and possibly 2021. Postpone doing this exercise until you are a bit more comfortable with handling functional data on R. Use data from mortality curves on Denmark (*`mortality.org`*), which has data until 2021.* □

  Exercise 2. *Think that you have, as well as male mortality curves, some female mortality curves on the same years. What would you do? What questions would you like to ask about the dataset of both curves? What would happen if you aggregate mortalities?*

  Imagine that the data, instead of being functional, is multivariate. What methods from multivariate statistics would you use, and what questions might those methods answer?

  Answer this question in prose, no need for fancy math or computation. This is just a thought experiment. □

- Finance (3 papers): For example, Bapna et al. (2008) uses FDA techniques to study the dynamics of prices in online auction markets. They study the first and second derivatives of the data, which would not be possible without FDA!

- Neurology (2 papers): For example, Buckner et al. (2004) uses normalization techniques from FDA to study brain images in demented patients.

- Economics (2 papers): For example, Ramsay and Ramsey (2002) consider differential equations estimated from functional data to study the dynamics of production indexes in a 70 year span.

- Engineering (2 papers): For example, Dabo-Niang et al. (2007) consider modal curves, as well as other central curves (median, mean) to propose a nonparametric clustering method for radar waveforms in the Amazonian basin.

- Agriculture (1 paper): For example, Ogden et al. (2002) transform images of fields to functional data, and uses that data to predict lodging score (a measure of how tilted a given crop is, if the crop is more tilted then it might lead to lowered yields or diseases on the crop) for rice fields via functional regression and functional principal component analysis.

- Physiology (1 paper): For example, Newell et al. (2006) uses FDA to compare lactate curves between athletes. Lactate curves are a way to measure athlete performance.

- Information Technology (1 paper): For example, Stewart et al. (2006) study the dynamics of open software evolution in the open source context. An interesting conclusion: software complexity tend to decrease as software grows in size.

- Education (1 paper): For example, Rupp (2005) clusters students taking a certain Mathematics and Science standardized exam based on sociodemographic and attitudinal variables. Once the groups are selected, curves for scores of groups are compared using functional principal components.
- Chemistry (1 paper): For example, Hutchinson et al. (2004) studies molecular weight distribution data (different weight fraction is expressed as a function of chain length, hence FDA) with functional regression models.
- Geophysics (1 paper): For example, Maslova et al. (2010) uses a functional regression model to test if auroral substorms and euqtorial and mid-latitude currents interact in such a way that is statistically significant.
- Behavioural science (1 paper): For example, Zhang et al. (2006) predicts fruit flies longevity using their sexual signalling curves. They conclude that high sexual signaling predict extended lifespan!

There are also more novel applications on the same and other areas, of course. For example, Patil et al. (2022) use functional principal components to reduce dimensionality of complex astrochemical data; our very own Henry Laniado has a Nature article (Azcorra et al., 2018) which uses unsupervised outlier detection techniques for finding influential users in social networks; your humble servant Calle-Saldarriaga et al. (2021) use a novel homogeneity test procedure for chemical data; Claeskens et al. (2014) propose a depth measure for multivariate functions and detect outliers on weather data; Zhang et al. (2016) study neural connectivity using generalized functional linear models; Delicado et al. (2010) study spatial functional data and proposes a functional Kriging method; Cuevas et al. (2004) proposes as ANOVA test for functional cardiology data; Fraiman and Muniz (2001) proposes a robust mean for functional data and applies it to NASDAQ data.

Exercise 3. *What is the topic of your thesis? Do a brief search of functional data applications on that domain, and select a few papers. Read the abstract and try to summarize the important ideas of the abstract. For example, let's say you are applying machine learning to analyze marketing trends in your thesis. So you do a quick search of functional data marketing and select some papers that interest you, and summarize their abstracts. If you do not have a topic for your thesis yet, don't worry. Just try to search for something you are interested on.* □

2.2 List of people

A (very biased) hall of Fame of Functional data analysis:

- Hans-Georg Muller: From UC Davis Statistics department, an eminence of the field. Has published on various applications (mostly for biological data), and many theoretical papers that have helped advance the area.
- Jane-Ling Wang: Also from UC Davis statistics department, also publishes many biological/medical applications of FDA, as well as sound theoretical papers. Publishes with Muller a lot. Head of UC Davis' research group on FDA.
- James O. Ramsay: From McGill University's Pschology department. Coined the term Functional Data Analysis, considered the father of the area (as well as Bernard Silverman). Publishes mostly applications, but many times they are so novel that they require new methodological advances. Author of many influentiall introductory textbooks on FDA.
- Bernard Silverman: Joint position at University of Bristol and Oxford. Publishes many interesting theoretical papers, joint author of the most influential FDA book with Ramsay.
- Ricardo Fraiman: From Universidad de la República de Uruguay, centro de Matemáticas. Has published many novel techniques for FDA, translating tools from multivariate data analysis to this novel world.
- Antonio Cuevas: From Universidad autónoma de madrid. Publishes constantly with Fraiman, has proposed many novel and widely used FDA techniques.

- Peter Hall: From Australia National University/University of Melbourne/UC Davis. Sadly he passed away in 2016. He is considered as one of the most influential and prolific statisticians of all time. His research areas span martingales and rates of convergence, extremes, coverage processes, bootstrapping, high dimensional data, deconvolution, and of course, functional data analysis. He has an staggering 606 publications.

  Exercise 4. *Peter Hall was so beloved and influential that he recieved not one, but two obituaries in great statistical journals: (Müller, 2016) and (Robinson and Welsh, 2018). Check out both papers to know more about his biography and interests. Pick two papers for functional data analysis, based on whatever criteria you have, and read the abstracts. Write out the FDA methodology they use in those papers.* □

- Frederic Ferraty (Toulouse University) and Philippe Vieu (Paul Sabatier university): Dynamic french duo, they published the authorative book on nonparametric FDA methods. Very interesting theoretical papers from both of them.

- Giles Hooker: From Cornell University, he researches FDA and machine learning, and bring many interesting insights from ML that many classical statistician don't normally consider.

- Stanislav Nagy: One for the future. Just started his academic career. From Karlova University. Has just started publishing in 2015, but since then has made monstrous contributions for FDA.

Here in Colombia, FDA is still in its early infancy. I attended the 2019 Colombian symposium in statistics, presenting a talk in FDA. Surprisingly, there was an FDA section, with like 8 or so talks. To be brutally honest, the research was not very good (including mine). We have a lot to do in order to grow in this area. The researcher with the most interesting work in FDA here is probably Ramón Giraldo, from Universidad Nacional in Bogotá. Most specifically, in EAFIT, we have some people interested in FDA: Henry Laniado, Francico Zuluaga, Nicolás Moreno, Santiago Ortiz, and me, Alejandro Calle Saldarriaga. Henry introduced all to the area[1] (me in 2017, and I have been amazed, confused and thrilled ever since) and it has been very nice researching in this novel and strange terrain. We are just beginning to do cool stuff, and have already some papers to show for it. More to come for sure! Sadly I will be leaving EAFIT shortly but I will for sure be still in touch with my friends and colleagues here.

2.3 Why FDA?

This is a question I got asked last time I presented this course. I had an answer (mostly appropriate I believe). But I don't want to answer this question. I want to show you the answer of this question indirectly, through the course.

Exercise 5. *Why do people use FDA instead of lets say, high-dimensional statistics or time series analysis? You are probably not in the capacity to answer this question right now. Try to answer it. You will refine your answer as the course progresses.* □

2.4 Methods in FDA

Almost any method you find for multivariate data analysis could be found also on functional data analysis, appropriately changed. A nice review of usual methodology used in FDA is presented in (Wang et al., 2016). You can also check the CRAN Task View for functional data analysis to check what methodologies are implemented in R (`https://cran.r-project.org/web/views/FunctionalData.html`).

Exercise 6. *Check the abstract and the sections of (Wang et al., 2016). Which methods intrigue you? Which would you want to learn? From the methods that interest you, pick two papers and try and summarize the abstract.* □

End of Lecture 1 material.

[1] except Nicolás, he came to Colombia with knowledge of the area I believe

3 Hands on FDA

3.1 Smoothness: Some mathematical considerations

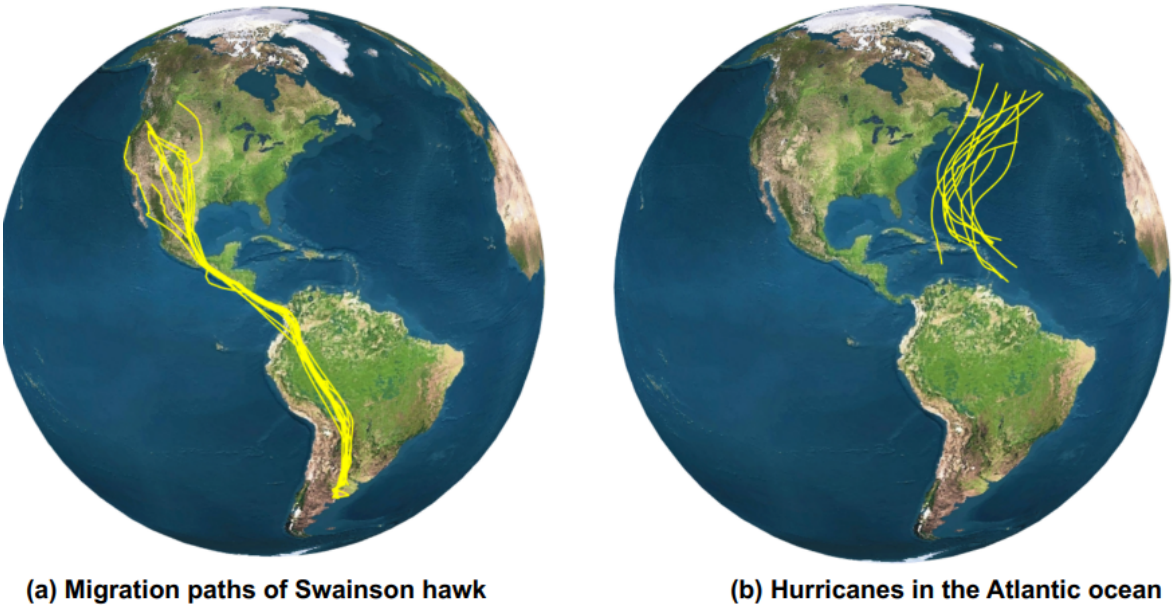
We are concerned, primarily, with univariate smooth functions: functions with one variable, one response, or more formally, functions $f : \mathbb{R} \to \mathbb{R}$, that is, real valued functions. These are the main object we will study in our course. However, that is not the only type of functional objects that can be considered.

¿Are all functions that arise in nature univariate? Of course not. For example, Górecki et al. (2018) present some methods for FDA on the multivariate domain. They propose a framework in which each individual is a country with the fourth following responses, which are indexed in time: GDP, Energy use, CO_2 emissions and population of urban agglomerations of more than 1 million as a percentage of the population.

¿But are all univariate functions that arise in nature real valued? Of course not. For example Calhoun and Adali (2012) propose that FMRI signals have a real and imaginary parts, and analyze them as such.

Other interesting functions can be of the type $f : \mathbb{R} \to \mathbb{Z}$. For example, covid cases in different countries per day. For example, (Kowal, 2019) uses integer-valued functional data for measles forecasting. But how do smoothness looks like here?

¿And is $\mathbb{R}$ the only topological space where functions can live? In other words, ¿can we propose another topological space, let's call it $\mathbb{T}$, such that we have smooth functions $f : \mathbb{R} \to \mathbb{T}$, i.e., we measure something trough time in that topological space? The answer is, of course, yes: manifold valued topological data analysis. For example, Zhang and Saparbayeva (2022) considers functions $f : \mathbb{R} \to \mathbb{S}^2$, which have many applications, see Figure 2 for examples of these kinds of functions.



(a) Migration paths of Swainson hawk

(b) Hurricanes in the Atlantic ocean

Figure 2: Manifold valued functional data.

But what exactly is a smooth function? Intuitively, a smooth function in an interval is a function for which you can compute derivatives in that interval. There are degrees of smoothness (in mathematics: there are many differentiability class): the more you can derive it while arriving at continuous functions, the smoother it is (or more precisely: it belongs to a higher differentiability class).

Definition. Let $\mathbb{I} \subseteq \mathbb{R}$ be an interval. Let $k \in \mathbb{Z}$, with $k \geq 0$. Then the function is from the differentiability class C^k if $f', f'', \ldots, f^{(k)}$ exist and are continuous in $\mathbb{I}$. The function is said to be of class C^∞ if it has derivatives of all orders in $\mathbb{I}$ that are continuous in $\mathbb{I}$. If f is of class C^∞ if is called smooth.

Exercise 7. *Let $B : \mathbb{R} \to \mathbb{R}$ with*

$$B(x) = \begin{cases} exp\left(-\frac{1}{1-x^2}\right) & \text{if } |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

Prove that this function is of class C^∞. Hint: Derivate two times, note the form of the derivatives. Note also that the product of two smooth functions is smooth. You have to show that $B'(-1) \to 0$ and $B'(1) \to 0$, and similar arguments for all higher derivatives.

□

3.2 Smoothing

Let's return to the realm of statistics. The simplest dataset encountered in FDA is of the form

$$x_n(t_{j,n}) \in \mathbb{R}; \quad t_{j,n} \in [T_1, T_2]; \quad n = 1, 2, \dots N; \quad j = 1, \dots, J_n$$

This means that we have N curves, observed on a common interval $[T_1, T_2]$, available at specific time points, $t_{j,n}$, which could be different for different curves. Let's assume that the number of points per curve is high: that might not be the case, but special care must be taken if it is not (Sparse FDA).

FDA studies, then, the set of curves

$$\{X_n(t) : t \in [T_1, T_2], n = 1, \dots, N\},$$

for which values exists at any $t \in [T_1, T_2]$, which are smooth. So how do we go from the first representation to the second?

We express the functional data by means of basis expansions:

$$X_n(t) \approx \sum_{k=1}^{K} c_{nk} B_k(t), \qquad 1 \le n \le N.$$

The basic intuition around this is that the smooth functions considered in a given sample share some shape properties, so each one can be approximated as a linear combination of some basic shapes, B_m. If the original points $t_{j,n}$ differ between individuals, the expansion put the curves into a common domain so that they are easily comparable. In the R package fda there are many basis functions.

But let's look at the problem more closely. Let's say we have a function

$$X(t) \approx \sum_{k=1}^{K} c_k B_k(t) = \mathbf{c}'\mathbf{B}$$

where $\mathbf{c}'$ is a vector of length K which contains the coefficients c_k and B is a $n \times k$ matrix containing the values $B_k(t_j)$.

We want $X(t)$ to be as close as $\sum_{k=1}^{K} c_k B_k(t)$ as possible, so an option is to minimize the sum of least squares

$$SSE = \sum_{j=1}^{N} [y_j - \sum_{k=1}^{K} c_k B_k(t_j)]^2$$

This is just a classic least squares problem! If you remember the formula for the regression coefficients in multiple linear regression, you should note that

$$\hat{\mathbf{c}} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'y$$

and our fitted values are then

$$\hat{y} = B\hat{\mathbf{c}}$$

Note that this is an OLS procedure, with all its problems and assumptions, and if we don't check those assumptions correctly we will have all the problems we have while fitting OLS, so you must be careful with this.

Exercise 8. *Read section 4.2.2 in (Ramsay et al., 2005). Implement a weighted least squares fit for the weather data we used in class, with the same number of base functions. Compute $\hat{y}_{UW}$ (unweighted fitted values), $\hat{y}_W$ (weighted fitted values). Compare $RMSE(y, \hat{y}_{UW})$ and $RMSE(y, \hat{y}_{UW})$ for each curve. Report means and standard deviations of these errors, conclude. Plot the two sets of smoothed curves.* □

3.3 Some common bases

Okay, now let's get our hands dirty and actually work with some functional data. The first thing we will do is smoothing, using the package FDA (Ramsay et al., 2021) in R. Of course, you can work with whatever language you like the most, but you'll have to search for packages which do the things I do in R, or, *shudders*, implement them yourself.

We then smooth functions by representing them as a sum in some base of functions, for example, the base of monomials

$$1, x, x^2, \ldots, x^k, \ldots$$

These functions are not very useful for approximating functions, as the leading term tends to be of much more magnitude than all the others. However, the Stone–Weierstrass theorem states that every continuous function on a given interval can be approximated as closely as desired by a given polynomial. Since polynomials are just linear combinations of monomials, that means we could approximate any function using the monomial basis. But in practice, this does not converge quickly enough.

The Fourier base is

$$1, sin(\omega t), cos(\omega t), sin(2\omega t), cos(2\omega t), \ldots$$

or it could also be expressed as

$$B_1(t) = 1; B_{2r-1}(t) = sin(r\omega t); B_{2r}(t) = cos(r\omega t)$$

where ω determines the period, $\frac{2\pi}{\omega}$. Note that

$$\frac{d}{dt} sinr\omega t = r\omega cosr\omega t$$
$$\frac{d}{dt} cosr\omega t = -r\omega sinr\omega t$$

so if the original basis of X has the coefficients $(c_0, c_1, c_2, c_3, \ldots, c_K)$ then the Fourier basis expansion of X' has the coefficients $(0, c_1, -\omega c_2, 2\omega c_3, -2\omega c_4, \ldots,)$.

Exercise 9. *Find the coefficients for the Fourier basis expansion of X'' and X''' in terms of the Fourier basis expansion coefficients of X.* □

The Fourier series is good for periodic data, which don't have strong local features and are stable. Ideally, the periodicity of the series should be reflected on the data.

Exercise 10. *Two functions defined on the same interval $\mathbb{I}$ are orthogonal if*

$$< x(t), y(t) >= \int_{\mathbb{I}} x(t)y(t)dt$$

Let $\mathbb{I} = [-2\pi, 2\pi]$. Prove that all the functions on the Fourier basis on $\mathbb{I}$ are orthogonal to every other function in the Fourier Basis. That means that the Basis is an orthogonal basis on $\mathbb{I}$. □

.

Another interesting basis is the B-spline basis. I won't get into all the details of what is an spline, but I will illustrate its concepts. Let's say you have a sampled function, i.e., you have a set of points of the form (x, y) which you want to approximate as a spline of order m. Let's say that $a \leq x \leq b$. First, divide $[a, b]$ into knot points $\tau_0, \tau_1, \ldots, \tau_L$. Then, a spline is a series of piecewise polynomials of order m defined between knots, which respect the values at the knots and possible some smoothness constraints at the knots.

A spline basis is then using a number of splines to approximate a given function. It is worth noting that a linear combination of splines is still a spline, so we are essentially approximating sample members as splines. The order m of the spline will tell us how many times we can differentiate it and still obtain informative derivatives.

Another interesting question is how to obtain a suitable number of basis functions K. This is related to the bias-variance decomposition of the MSE: if you select few basis functions, your approximation will be very biased but with small variance. If you select many basis functions, your approximation will have very little bias but a high variance. An interesting algorithm to choose the number of basis, via crossvalidation, for a given curve is:

1. Leave out one observation (t_j, Y_j).
2. Fit the remaing data and get $\hat{X}_{-j}(t_j)$ for many k.
3. Choose k that minimizes $CV(X) = \sum_{j=1}^{m}(Y_j - \hat{X}_j(t_j))^2$

Exercise 11. *Implement the above algorithm. Run it for the precipitation dataset we considered in class. Check the number K of selected basis functions for each curve. How variable is it? What is the mean, median, mode? Draw an histogram. Conclude.* □

Now, an example. Let's consider the Montreal daily temperature database, which consists on daily measures of temperature for many years in the Canadian city. The unsmoothed data is plotted in Figure 3

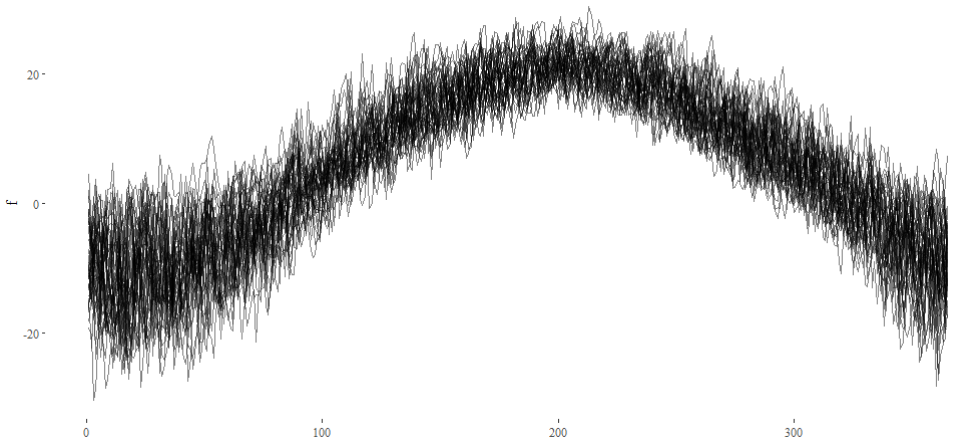


Figure 3: Montreal daily weather, unsmoothed

Now, let's use the package FDA to smooth using a Fourier basis. The code I use is:

```
library(fda)

TempDaily <-t(MontrealTemp)
basis <- create.fourier.basis(c(1,365), 65)
range <- 1:365
basismat <- eval.basis(range, basis)
coefs <- lsfit(basismat, TempDaily,
```

```
                intercept=FALSE)$coef

tempfd <- fd(coefs, basis)
plot(tempfd)
temp_mean <- mean.fd(tempfd)
lines(temp_mean, lwd=2.5, col='red')
temp_sd <- sd.fd(tempfd)
lines(temp_sd, lwd=2.5, col = 'cyan')
```

The code, explained:

1. Load fda, load data.
2. Create a Fourier basis with 65 functions in the range $(1, 365)$.
3. Evaluate the basis in the range to obtain matrix $\mathbf{B}$.
4. Solve the least squares problem which gives us the coeffients of the basis, $\mathbf{c}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y}$, where $\mathbf{y}$ is the original data. This is done for each curve.
5. Create a FD object with the estimated coefficients and the defined basis.
6. Plot the smoothed data.
7. Estimate the mean function of the data and plot it.
8. Estimate the standard deviation of the data and plot it.

A helper function for doing some of the things we are doing here is the `smooth.basis` function. An example of smoothing another dataset, using the b-spline basis is:

```
library(fda)
library(fda.usc)

data(tecator)

range <- tecator$absorp.fdata$argvals
y <- t(tecator$absorp.fdata$data)

splinebasis <- create.bspline.basis(tecator$absorp.fdata$rangeval, 15)

fd_data <- smooth.basis(range, y, splinebasis)

plot(fd_data)
```

Exercise 12. *The tecator dataset has some other covariates. Separate the curves in two types of curves: Curves which have a high fat percentage (*`tecator$y$Fat` ≥ 20*), plot them in the same plot but with different colors. Can you note differences?* □

Exercise 13. *One cool feature of functional data is that we can examine the derivatives of our functions. Read the documentation for the* `deriv.fd` *function. Compute the first and second derivatives of the data (after smoothing). Plot the derivatives for curves with low and high fat with different colors. Can you note differences in the groups?* □

3.4 Bibliographic note

For an introduction to basis functions, you can read section 3 of (Ramsay et al., 2005). For an introduction to smoothing you can read section 4 of (Ramsay et al., 2005). The Wikipedia page for smoothness (in the mathematical sense) is a good introduction `https://en.wikipedia.org/wiki/Smoothness`. For a complete introduction and implementation of B-splines refer to (d. Boor, 2001).

End of lecture 2.

4 More exploratory tools

4.1 Some common statistics

Let's assume now that we are using the fully functional representation of our objects, i.e., we have successfully smoothed the data, which is now of the form

$$\{x_n(t) \quad , t \in [T_1, T_2] \quad , n = 1, \ldots, N\}$$

The simplest summary statistics are

$$\bar{x}_N(t) = \frac{1}{n} \sum_{n=1}^{N} x_n(t)$$

$$sd_X(t) = \left\{ \frac{1}{N-1} \sum_{n=1}^{N} (x_n(t) - \bar{x}(t))^2 \right\}^{1/2}$$

which are the pointwise mean and standard deviation.

Sample standard deviation gives us an idea of variability in point t. But what if we wanted to know hot the variability at point t relates to the variability of point s? We can consider the sample covariance function

$$\hat{c}_X(s,t) = \frac{1}{N-1} \sum_{n=1}^{N} (x_n(s) - \bar{x}(s))(x_n(t) - \bar{x}(t))$$

these are almost indistinguishable from their univariate or multivariate counterparts. The key difference is that instead of having a mean or variance value (or vector), we have a mean or variance *function*, and instead of having a Covariance matrix, we have a covariance function of two variables. Correlation can also be defined in a similar manner. Some higher order moments can be defined for fda, see for example (Staicu et al., 2012) for skewness in fda, and (Vidal et al., 2021) for kurtosis.

Exercise 14. *Use the rainbow plot, from package* `rainbow` *in R, to visualize the functional boxplot of both datasets showed in class until this point (temperature data for Canada, el niño/niña temperature in Oceans) with function* `fboxplot`*. Please note that the functional data object used in this package is different than the one used in package* `fda`*. Format accordingly. Now, detect outliers in both datasets, using the functions* `foutliers` *from* `rainbow`*. If something is not clear about the usage, read the documentation of the package (*`https://cran.r-project.org/web/packages/rainbow/rainbow.pdf`*) and the original paper (Hyndman and Shang, 2010).* □

Exercise 15. *We have expressed functions $x_n(t)$ as $x_n(t) \approx \sum_{m=1}^{M} c_{nm} B_m(t)$. One can show that the mean can also be expressed as $\bar{x}_n(t) = \sum_{i=1}^{M} a_m B_m(t)$ and the covariance as $\hat{c}(t,s) = \sum_{m=1}^{M} \sum_{m=1}^{M} b_{mk} B_m(t) B_k(s)$. Find the expressions for a_m and b_{mk} assuming that $B_m(t)$ are an orthonormal basis. An orthonormal basis is an orthogonal basis with the additional condition that the norm of the functions is 1.* □

Now, an additional feature of fda than multivariate data analysis does not have is the ability to take derivatives of the data. If we have a basis representation, you can just define the k-th derivative as

$$x_n^{(k)}(t) \approx \sum_{m=1}^{M} c_{nm} B^{(k)}(t)$$

Note that the amount of informative derivatives we obtain from the data is dependent on the form of the original basis functions.

4.2 Penalized smoothing

The method studied until now for smoothing work when the original data is already somewhat smooth. However, if the original data contains a high level of noise, the functional objects constructed in this way will inherit this variability, resulting in very wiggly, variable curves. An example of this is precipitation data: let's say we have daily observations of precipitation data in a city. For example, precipitation in summer is very high, approximately, 50 mm daily average. Let's say we have two consecutive days, one with 58 mm of precipitation, and one where it does not rain. Would we like for a curve to be close to those two points?

In order to solve this problem, we use something called penalized smoothing. For a given curve, suppose we observe data (t_j, y_j). We assume that there exists a smooth curve $x(t), \quad t \in [t_1, t_2]$, and $y_j = x(t_j) + \epsilon_j$, where $E[\epsilon_j] = 0$. The goal of smoothing is to try and elimate the contribution of ϵ_j. Once again, we will get the representation

$$x(t) \approx x_K(t) = \sum_{k=1}^{K} c_k B_k(t)$$

where K is often much larger than M, and can be even bigger than the number of observations J. More parameters than observations? Don't worry, penalization deals with that!

Let L be a linear differential operator (we will be covering linear operators more closely later, but just think of linear differential operators as linear combinations of derivatives).

$$L[x(t)] = \alpha_0(t)x(t) + \alpha_1(t)x'(t) + \cdots + \alpha_m(t)x^{(m)}(t)$$

The optimization problem to solve now is

$$PSS_\lambda(c_1, \ldots, c_k) = \sum_j (y_j - x_K(t_j))^2 + \lambda \int_{T_1}^{T_2} [L(x_K(t))]^2$$

where λ is called the smoothing parameters. If λ is 0, we will get usual smoothing. If λ is too large, the second term becomes becomes dominant, and the observed data may not be too close to the smoothed curves. A good λ strikes a balance between overfitting and oversmoothing. We will use automated methods to choose it, but a good idea is to plot it and see if they reflect data in an appropriate way.

A good and widely used operator for periodic data (Fourier basis) is called the harmonic accelerator operator. Suppose you observe $x(t)$ at t_j points, with $j = 1, \ldots, J$. Let $K = 2J + 1$, $T_2 - T_1 = T$. Assume $T_1 = 0$ for simplicity. then $x(t)$ can be expressed as

$$x(t) = c_0 + \sum_{j=1}^{J} [a_j sin(\omega j t) + b_j cos(\omega j t)], \quad \omega = \frac{2\pi}{T},$$

then the harmonic accelerator is

$$L[x(t)] = \omega^2 x'(t) + x^{(3)}(t)$$

Exercise 16. *Show that, for the harmonic accelerator,*

$$\int_0^T [L(x(t)]^2 dt = \pi\omega^5 \sum_{j=2}^{J} j^2(j^2 - 1)^2(a_j^2 + b_j^2)$$

□

Exercise 17. *Federal reserve yields are yield curves that are used to measure interest rates with respect to maturity. You can find a dataset of Federal Reserve Yields in* 1982 *in the R package* `fds` *(Shang and Hyndman, 2018).*

Smooth the curves using B-splines with 4 *basis functions. Plot raw and smoothed interest rates on the same plot. Penalize using the second derivative as a a penalty operator, finding a suitable* λ*. Plot the raw and the penalized smooth data. Which do you prefer?* □

4.3 Functional Principal Components

Principal components have also been generalized from multivariate statistics to the functional setting, and are one of the most useful and often used tools in FDA (and, in my opinion, more natural). Estimated FPC (functional principal component) are related to the eigenfunctions of the covariance function, $\hat{c}(t,s)$, akin to multivariate statistics, where the covariance matrix determines the principal components.

The idea of FPCA is that, having found a smoothing for the data, with M basis functions, to find the eigenfunctions $\hat{v}_j$ such that

$$X_n - \bar{X}_N(t) \approx \sum_{j=1}^{p} \hat{\xi}_{n,j}\hat{v}_j(t),$$

where each $\hat{v}_j$ represents the j-th most pronunced pattern of deviation from the mean sample function, and the coefficients $\hat{\xi}_{n,j}$ quantifies the contribution of each FPC to the shape of $x_n(t)$, and is called the score of $x_n(t)$ with respect to $\hat{v}_j(t)$.

The $\hat{v}_j(t)$ hold this interesting property:

$$\int \hat{v}_j\hat{v}_i dt = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if} i = j \end{cases}$$

and is also a basis in the space considered. That means that the eigenfunctions of the covariance function form an orthonormal basis for the space of functions we are considering. The percentage of variability explained by $\hat{v}_j$ is related to the size of the scores $\hat{\xi}_{n,j}$. If you consider the eigenvalues, in order descending order $\hat{\lambda}_1, \hat{\lambda}_2, \dots$ you can find the percentage of explained variance as

$$CPV(m) = \frac{\sum_{k=1}^{K} \hat{\lambda}_k}{\sum_{k=1}^{N} \hat{\lambda}_k}$$

But in order to understand how to estimate $\hat{v}_j$ and $\hat{\lambda}_j$, we need some more mathematical concepts, which we will get into later. Now we will just use FPC as an exploratory tool and will estimate the functions with some tools already implemented in R.

Exercise 18. *The pinch dataset is included in the* `fda` *package. It consists of* 151 *measurements of pinch force for* 20 *curves. Convert the data to functional data using* 15 *B-spline basis function or order* 4 *(cubic). Plot them. Compute the mean and standard deviation functions, add them to the plot. Graph perspective and contour plots for the sample covariance. Interpret. Graph the first four FPC of the data. How many components do you need to explain the* 90% *of the variance?* □

Exercise 19. *The Matérn covariance function leads to a very general family of stationary Gaussian processes. The function is*

$$\frac{\sigma^2}{\Gamma(v)2^{\nu-1}} \left(\frac{\sqrt{2\nu}|t-s|}{\rho} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|t-s|}{\rho} \right)$$

for $\nu > 0$*, where* σ^2 *is the variance parameter,* ν *is the smoothness parameter, and* ρ *is the range parameter.* $K_\nu(\cdot)$ *is a modified Bessell function of the second kind. The paths of the Matérn process are* k *times continuously differentiable for any* $v > k$ *with probability* 1*.*

Simulate 100 *i.i.d Matérn processes with mean zero, for* $\nu = 1/2, 2, 4$*. Set* $\sigma^2 = 1$ *and* $\rho = 1$*. Use a temporal grid of* 50 *evenly spaced points in* $[0,1]$*, and evaluate* $c(t,s)$ *in this grid to get a covariance matrix of* 50×50*, call it* Σ*.*

To evaluate K_ν*, use the R function* `BesselK(x, nu)` *where* `x` *is the value* $\frac{\sqrt{2\nu}|t-s|}{\rho}$ *and* `nu` *is* ν*. Now, simulate each trajectory using the covariance matrix* Σ *and a mean vector filled with zeroes, using a multivariate random vector generator.*

For each $\nu = 1/2, 2, 4$*, plot the simulated trajectories. Plot the first four FPC, comment on similarities/differences. Plot the standard deviation for each* ν*, comment on similarities/differences. Create a plot for the covariance structure for each* ν*, comment on similarities/differences.* □

4.4 Bibliographic note

For a more rigorous definition of the statistics we considered, see section 2.4 of (Horváth and Kokoszka, 2012). FDA has also been applied to smooth solutions to differential equations, see (Ramsay et al., 2007). Another interesting method for exploratory data analysis on curves is called curve alignment, which considers that different curves in the same sample are similar to one another except that they are shifted in some way, and that it is of interest to align them. See (Marron et al., 2015) for a thorough review.

End lecture 3.

5 Some mathematical prerequisites

5.1 The space of functions

To simplify notation, lets assume all functions we consider are in $[0,1]$. This does not limit our expressive power: consider a function f in $[a,b]$. Then we can simply define the function f^*, such that $f^*(u) = f(t)$ where $u = \frac{t-a}{b-a}$, which retains the shape properties an is defined in $[0,1]$.

A function is said to be square-integrable if

$$\int f^2(t)dt = \int_0^1 \{f(t)\}^2 dt < \infty.$$

If the limits of integration are missing, assume that we are integrating over $[0,1]$. Let's call the space of all functions that are square-integrable L^2.

Square integrable functions form a vector space. That means that if $f, g \in L^2$, then $af + bg \in L^2$ for some scalars a, b.

In L^2 we can define the inner product

$$< f, g >= \int f(t)g(t)dt.$$

We say that functions are orthogonal if

$$< f, g >= 0.$$

This allows us to introduce the notion of distance between functions via the norm

$$||f|| = \sqrt{< f, f >} = \left\{ \int f^2(t)dt \right\},$$

so that the distance between functions in L^2 is defined as

$$d(f, g) = ||f - g||.$$

Properties of an inner product space. A vector space $\mathbb{V}$ is an inner product space if for each $x, y \in \mathbb{V}$ there is a scalar $< x, y >$, called the inner product, which has this properties:

i) $< x, y >=< y, x >$

ii) $< a_1x_1 + a_2x_2, y >= a_1 < x_1, y > +a_2 < x_2, y >$

iii) $< x, x >\geq 0$, and $< x, x >= 0$ if and only if $x = 0$.

Exercise 20. *Prove that L^2 is an inner product space.* □

Properties of the norm. Let $|| \cdot ||$ be a norm defined by the inner product. Then

i) $||ax|| = |a|||x||$.

ii) $| < x, y > | \leq ||x||||y||$, called the Cauchy-Schwartz inequality.

iii) $||x + y|| \leq ||x|| + ||y||$, called the triangle inequality.

iv) $d(x, y) = ||x - y||$ is a metric.

Exercise 21. *Prove that i, ii, iii hold for the norm we defined in L^2.* □

As we have seen, basis expansions play a role in fda. We say that a set of functions $\{e_1, e_2, \dots\}$ is a basis in L^2 if every $f \in L^2$ admits a unique expansion

$$f(t) = \sum_{j=1}^{\infty} a_j e_j(t)$$

If $\{e_1, e_2, \dots\}$ is an orthonormal basis, we have that $a_j =< f, e_j >$ and we have Parseval's equality

$$\int f^2(t)dt = ||f||^2 = \sum_{j=1}^{\infty} < f, e_j >^2 = \sum_{j=1}^{\infty} \left\{ \int f(t)e_j(t)dt \right\}^2$$

5.2 Random elements

Like in usual statistics, we have samples $x_1, \ldots, x_N$, which we say are realizations of $X_1, \ldots, X_N$. In univarirate statistics, we say each $X_1, \ldots, X_N$ is a random variable, and in multivariate statistics we say it is a random vector. In the fda case, we say they are random functions.

Just as random variables, random functions are defined in a probability space, Ω, such that for each $\omega \in \Omega$, $X(\omega)$ is a determinsitic function. We assume all realizations $X(\omega)$ are elements of L^2, that is

$$||X(\omega)||^2 = \int \{X(\omega)(t)\}^2 < \infty$$

The function $\omega \to ||X(\omega)||$ is a classic random univariate variable, and we can ask things like $E||X||^2 < \infty$, that means, if it has a finite second moment (or in this case, if it is square integrable). For a random function X, we define

$$\mu(t) = E[X(t)], \qquad c(t,s) = E[(X(t) - \mu(t))(X(s) - \mu(t))]$$

where $\bar{X}_N(t)$ and $\hat{c}(t,s)$ we considered earlier are estimators of μ and c.

By the Karhunen-Loéve, we can express any square integrable function X as

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j v_j(t)$$

where the v_j depend on the covariance function c of X. Specifically, they are the eigenfunctions of c, i.e., they are the solutions to the equation

$$\int c(t,s)v(s)ds = \lambda v(t)$$

there are countable many solutions, $(\lambda_1, v_1), (\lambda_2, v_2), \ldots$, which are called the eigenvalue-eigenfunctions pairs of X (or c), such that they are arranged in nonincreasing order $\lambda_1 \geq \lambda_2 \geq \ldots$. The random variables ξ_i, called the scores, are

$$\xi_j =< X - \mu, v_j >= \int (X(t) - \mu(t))v_j(t)dt$$

and it can bee shown that

$$E[\xi_j] = 0, \qquad E[\xi_j^2] = \lambda_j, \qquad Cov[\xi_j, \xi_k] = 0 \qquad \text{if } j \neq k$$

and

$$E\left[\int (X(t) - \mu(t))^2 dt\right] = E||X - \mu||^2 = \sum_{j=1}^{\infty} \lambda_j$$

So the λ represent the variance of X. Functions v_j form an orthonormal basis, and are called the FPC of X. They are optimal in the sense that X can be approximated by a few of them.

Exercise 22. *Consider a standardized ($\sigma^2 = 1$) Brownian Motion Process, or Wiener process in an interval $(0, T)$, and call it $W(t)$. The Brownian bridge is defined as*

$$B(t) = W(t) - tW(T)$$

The mean and covariance of this process are:

$$\mu(t) = 0$$
$$c(s,t) = s(1-t)$$

The eigenfunctions/eigenvalues pair for this process are

$$v_k(t) = \sqrt{2} sin(k\pi t)$$
$$\lambda_k = \frac{1}{k^2\pi^2}$$

Verify that the above are eigenfunction/eigenvalue pairs. Verify that the eigenfunctions are orthogonal. □

Exercise 23. *Let* $\{h_1(t), h_2(t), \dots\}$ *be a basis. One can define the following orthonormalization*

$$e_0 = \frac{h_0}{||h_0||}, \quad e_{n+1} = \alpha_n e_n + \beta_{n+1} h_{n+1}$$

where $\alpha_n = -\beta_{n+1} < e_n, h_{n+1} >$ *and*

$$\beta_{n+1}^2 = \frac{1}{||h_{n+1}||^2 - | < e_n, h_{n+1} > |^2}$$

This is called the Gram-Schmidt orthonormalization. Consider the first six functions in the base of monomials, x^m *with* $m = 0, 1, \dots$ *in* $[0, 1]$*. Plot the original elements, plot the orthonormalized elements.*

□

End of Lecture 4.

5.3 Linear Operators

Linear operators are defined in vector spaces. If $\mathbb{V}_1$, $\mathbb{V}_2$ are two vectors spaces, a function $L : \mathbb{V}_1 \to \mathbb{V}_2$ is called Linear if $L(ax + by) = aL(x) + bL(y)$ where $x, y \in \mathbb{V}_1$ and $L(ax + by) \in \mathbb{V}_2$.

In functional data analysis, we get this types of operators a lot of time

$$L_1(x) : L^2 \to \mathbb{R},$$

$$L_1(x) = \int \psi(t)x(t)dt$$

and

$$L_2(x) : L^2 \to L^2,$$

$$L_2(x) = \int \psi(t, s)x(s)ds$$

which are both called integral operators. L_1 transforms a function x to a real number $L_1(x)$, and L_2 transforms a function x to another function $L_2(x)$. It is easy to check linearity.

To ensure that the function $L_2(x)$ is square integrable, we must assume that

$$\int \int \psi^2(t, s)dtds < \infty \tag{1}$$

The functions ψ are called the Kernels of the respective operators, and an integral operator with Kernel satisfying Equation 1 is called a Hilbert-Schmidt operator.

An operator is called symmetric if

$$< L(x), y >=< x, L(y) >$$

Exercise 24. *Check if the following operators are symmetric*

$$i)L_1(x) = \int_0^t x(u)du$$

$$ii)L_2(x) = x(1 - t)$$

$$iii)L_3(x) = \int x(t)y(t - u)du$$

where $x, y \in L^2[0, 1]$ □

Now, an operator is called semi-positive definite if

$$< L(x), x >\geq 0$$

An operator that is used very often in FDA is called the covariance operator. The covariance function we discussed earlier $c(t, s)$ is the Kernel of this covariance operator. The operator is

$$C(x) = \int c(t, s)x(s)ds$$

The FPC are then the eigenfunctions

$$C(v_j) = \lambda_j v_j$$

The operator is symmetric since

$$\begin{aligned} c(t,s) &= E[X(t)X(s)] \\ &= E[X(s)X(t)] \\ &= c(s,t) \end{aligned}$$

and it is positive definite since

$$\begin{aligned} \int_0^1 \int_0^1 c(t,s)y(t)y(s)dtds &= \int_0^1 \int_0^1 E[X(t)X(s)]y(t)y(s)dtds \\ &= \int_0^1 E[X(t)]y(t)dt \int_0^1 E[X(s)]y(s)ds \\ &= \int_0^1 (E[X(t)y(t)]dt)^2 \\ &= E\left[(\int_0^1 X(t)y(t)dt)^2\right] \\ &\geq 0 \end{aligned}$$

where the last step follows from $x, y \in L^2$. This operator is also Hilbert-Schmidt. Since it is symmetric, positive definite, and Hilbert-Schmidt, it has the following convenient representation

$$C(x) = \sum_{j=1}^{\infty} \lambda_j < x, v_j > v_j$$

5.4 Bibliographic remarks

The mathematical theory for FDA can be much more rigorous and involved. See chapter 10, 11, 12 from (Kokoszka and Reimherr, 2017), or (Hsing and Eubank, 2015) for a deeper dive. The book by Debnath and Mikusinski (2005) can be used as a reference for functional analysis and Hilbert spaces, and the book (Akhiezer and Glazman, 2013) can be used as a reference for linear operators.

6 The functional Linear Model

The linear model. The holy grail of the first year statistics courses, usually introduced at the end of an univariate statistics sequence of courses or at the middle/end of multivariate statistics courses. Let's recall the linear model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i, \quad i = 1, \ldots, N \tag{2}$$

or in matrix notation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y}$ is a $N \times 1$ vector of responses, $\boldsymbol{X}$ is a $(N+1) \times p$ regression matrix of covariates ($+1$ because of the intercept), $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector and $\boldsymbol{\epsilon}$ is a $N \times 1$ vector of zero mean errors. Recall that by using least squares we can get the estimation $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}X^TY$, with plentiful of known diagnostics/hypothesis testing scenarios to be drawn from here. ¿So how do we turn this into its functional analogue? They are, actually, plenty of options to touch, we will cover here:

- ¿What if we turn responses y, which are usually scalars, into responses $y(t)$, which are functions? That is, we want to predict/explain functions using multivariate covariates. This is called function-on-scalar linear regression.

- ¿What if we keep responses as scalars, y, but want to use functional, $X(t)$ covariates? That is, we would want to predict/explain scalars using functions. This is called scalar-on-function linear regression.
- ¿What if we turn responses and covariates into functions? That is, we want to predict/explain functions using other functions. This is called function-on-function linear regression.

6.1 Scalar-on-function linear regression

The scalar-on-function linear regression with one functional predictor can be expressed as

$$y_i = \alpha + \int \beta(s)x_i(s) + \epsilon_i$$

where we want to estimate the scalar intercept β_0 which is the mean of y_i when $\beta(t) = 0$, and the functional parameter $\beta(t)$. There are many ways to estimate $\beta(t)$. The easier is by basis expansions. Assume that you can express the regression coefficient as a sum of basis functions

$$\beta(t) = \sum_{k=1}^{K} c_k B_k(t).$$

The selected basis functions B_k influence the shape of the estimate. The number K of selected basis functions will affect the smoothness of $\beta(t)$, and is usually chosen to be smaller than the number of points t_j at which the original covariates where observed. Choices of K can be data driven, as we will discuss shortly. Using the expansion

$$\begin{aligned}\int \beta(t)x_i(t)dt &= \sum_{k=1}^{K} c_k \int B_k(t)x_i(t)dt \\ &=: \sum_{k=1}^{K} c_k x_{ik}\end{aligned}$$

So we can express this problem as a problem in the form of Model 2

$$\begin{aligned}Y_i &= \beta_0 + \int \beta(s)x_i(s)ds + \epsilon_i \\ &= \beta_0 + \sum_{k=1}^{K} c_k x_{ik} + \epsilon_i\end{aligned}$$

which can be solved by usual least squares methods. The resulting estimate is

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{c}_k B_k(t)$$

where

$$\hat{c} = (X^T X)^{-1} X^T Y$$

and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & \dots & x_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{NK} \end{pmatrix}$$

The disadvantage is that the estimate depends on the shape of the chosen basis functions and the number of basis functions, so we get restricted shapes for $\beta(t)$, somewhat subjective. An usual approach is to use the same basis functions used for the basis expansion of the regressor curves $x_i(t)$. This estimate is biased because of the truncation done by approximating $\beta(t)$ with a finite number of basis functions.

Another method of estimating $\beta(t)$ is with a roughness penalty (to ensure smoothness and low variability, as shown in previous classes). In the precious estimation procedure, we had to choose K, which can be seen as a tuning parameter which adjusts the smoothness of $\hat{\beta}(t)$, but from this perspective, it is usually more desirable to smooth by using a roughness penalty term instead of tuning via K. For this method, we still have the expansion

$$\beta(t) = \sum_{k=1}^{K} c_k B_k(t)$$

where K is usually bigger and equal to the amount of points t, at which X_i is observed. The roughness penalty approach then solves the optimization problem

$$P_\lambda(\alpha, \beta) = \sum_{i=1}^{N} \left\{ Y_i - \alpha - \int \beta(t) X_i(t) dt \right\}^2 + \lambda \int [L(\beta(t))]^2 dt$$

where L is a linear differential operator applied on $\beta(t)$ and λ is the smoothing parameter. The idea is to penalize rough functions using the second term. A common choice for L is $L[\beta(t)] = \beta''(t)$. Once again, choosing λ is crucial. If it is too large, β is too smooth and important aspects of the curve may be suppressed.

When we use REML (restricted maximum likelihood) as the estimation procedure for getting $\hat{\beta}(t)$ we obtain estimations of random effect variance, and in turn this variances serve as the tuning parameter that controls the smoothness of $\beta(t)$ so we don't have to bother into choosing an appropriate λ.

A final way to estimate $\beta(t)$ is using functional principal components. First, remember that each $X \in L^2$ can be expanded as

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j v_j(t)$$

where v_j are the the eigenfuntions of the covariance operator. We can arrive at the standard linear regression model (Equation 2) by using the approximation

$$X_i(t) = \hat{\mu}(t) + \sum_{k=1}^{p} \hat{\xi}_{ij} v_j(t)$$

where

$$\hat{\xi_{ij}}(t) = \int [X_i(t) - \hat{\mu}(t)] \hat{v}_j(t) dt$$

which gives us the model

$$Y_i = \alpha + \int \beta(t)(\hat{\mu}(t) + \sum_{j=1}^{p} \hat{\xi}_{ij} \hat{v}_j) dt + \epsilon_i \tag{3}$$

$$= \alpha + \sum_{j=1}^{p} \hat{\xi}_{ij} \beta_j + \epsilon_i \tag{4}$$

where $\beta_0 = \alpha + \int \beta(t)\hat{\mu}(t)dt$ and $\beta_j = \int \beta(t)\hat{v}_j(t)dt$, which are the unkown parameters to estimate. Defining the $N \times (p+1)$ matrix

$$\Xi = \begin{pmatrix} 1 & \hat{\xi}_{11} & \hat{\xi}_{12} & \cdots & \hat{\xi}_{1K} \\ 1 & \hat{\xi}_{21} & \hat{\xi}_{22} & \cdots & \hat{\xi}_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \hat{\xi}_{N1} & \hat{\xi}_{N2} & \cdots & \hat{\xi}_{NK} \end{pmatrix}$$

we can estimate the parameter vector $\beta = [\beta_0, \beta_1, \ldots, \beta_p]^T$ by the least squares estimator of $Y = \Xi\beta + \epsilon$. Call $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ the estimatimes of this least squares problem. Then the functional parameters are

$$\hat{\beta}(t) = \sum_{j=1}^{p} \hat{\beta}_j \hat{v}_j$$

$$\hat{\alpha} = \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j \int \hat{v}_j(t) \hat{mu}(t) dt$$

An important aspect is to select the number p of estimate functional principal components. A simple approach is select the number of FPC which explain $85, 90, 95, 99$ percent of cumulative variance.

We have considered three different estimation methods. ¿Which is the best one? Goldsmith and Scheipl (2014) compare various estimation methods on many different datasets, and conclude that the choice of method depends on the data.

Exercise 25. *Consider the tecator dataset in* `fda.usc` *which is comprised of* $N = 125$ *meat samples, where the explanatory curves are spectral curves at some wavelengths and we have three possible responses: percentage of fat, percentage of water, and percentage of protein.*

i) For the 3 *possible responses, plot scatters for each pair and try to find relations between them.*

ii) Find the median of the three responses. A curve will be considered high if it has a higher response than the median, low if it does not. Plot on the same grid high and low curves (with different colours per group) for each response.

iii) Smooth the curves using penalized splines, finding an adequate λ *using methods discussed on class.*

iv) Derive the curves, repeat ii) with the derivatives of the curves. Comment which responses for derivatives or original curves where the groups more easily identifiable.

v) Plot smoothed functions, their mean, their standard deviation, analyze.

vi) Find FPCA adequately for the curves.

vii) Perform scalar on fuction regression using the three methods discussed. Plot $\hat{\beta}(t)$ *for all methods in the same grid. Compare* R^2*. Discuss.* □

6.2 Bibliographic remarks

For more information on functional linear models, refer to (Ramsay et al., 2005) where many chapters are reserved for this topic. You can also consider nonlinear scalar-on-function regression. One such method is based on functional generalized additive models, that take the form

$$Y_i = \alpha + \int f(X_i(t), t) dt + \epsilon_i$$

where f is smooth. See (McLean et al., 2014) for more information. This method is implemented as `af()` on `refund`. Also, we can consider the problem

$$Y_i = m(x_i) + \epsilon$$

where $m : L^2 \to \mathbb{R}$ is a functional that must be estimated. Ferraty and Vieu (2006) dedicate almost a whole book on such problem. The paper by Reiss et al. (2017) reviews many approaches for the scalar-on-function regression problem.

End lecture five.

6.3 Functional response models

The function-on-regression model can be written as

$$Y_i(t) = x_{i1}\beta_1(t) + x_{i2}\beta_2(t) + \cdots + x_{iq}\beta_q(t) + \epsilon_i(t), \qquad i = 1, \ldots, N \tag{5}$$

that means that each unit i you have a functional response Y_i with q scalar explanatory variables. The functions $\beta_1(t), \ldots, \beta_q(t)$ are generally called the effect functions. We can define the following matrix and vectors:

$$\boldsymbol{Y}(t) = \begin{pmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_N(t) \end{pmatrix}$$

,

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1q} \\ x_{21} & x_{22} & \ldots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Nq} \end{pmatrix}$$

,

$$\boldsymbol{\beta}(t) = \begin{pmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_q(t) \end{pmatrix}$$

,

$$\boldsymbol{\epsilon}(t) = \begin{pmatrix} \epsilon_1(t) \\ \epsilon_2(t) \\ \vdots \\ \epsilon_N(t) \end{pmatrix}$$

,

and then the model 5 can be written as

$$\boldsymbol{Y}(t) = \boldsymbol{X}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t),$$

so for any fixed t, you have an usual least squares problem, that could be solved by

$$\hat{\boldsymbol{\beta}}(t) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y}(t)$$

this assumes that $Y_i(t)$ are fully observed. But even if $Y_i(t)$ are observed at different time points for different curves we can first smooth and interpolate the missing values. If they are not observed in different times, $\hat{\beta}(t)$ can be estimated with the raw observations $Y_i(t_j)$.

Another possibly interesting analysis is to consider the residual curves

$$\hat{\boldsymbol{\epsilon}}(t) = \boldsymbol{Y}(t) - \hat{\boldsymbol{Y}}(t)$$

where

$$\hat{\boldsymbol{Y}}(t) = \boldsymbol{X}\hat{\boldsymbol{\beta}}(t)$$

and find its functional principal components. The way we could understand better the variance that was not explained by our initial function-on-scalar model. But we just estimated β_l pointwise, which is only justified if the $Y_i(t)$ are smooth enough. If this assumption is not met, we of course could use some penalties to ensure the smoothness of our effect curves. Assume that the effect funcions admit the expansion

$$\beta_l(t_j) = \sum_{k=1}^{K} b_{lk}\phi_k(t_j)$$

where $\phi_1, \phi_2, \ldots, \phi_K$ are some basis functions, which number (K) should be large. Assuming all responses are available for each t_j, $1 \leq j \leq J$, model in Equation 5 could be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}\boldsymbol{\Phi} + \boldsymbol{\epsilon}$$

with the matrices

$$\begin{aligned}
\boldsymbol{Y} &= [Y_i(t_j), \quad 1 \leq i \leq N, \quad 1 \leq j \leq N], \qquad (N \times J) \\
\boldsymbol{X} &= [x_{il}, \quad 1 \leq i \leq N, \quad 1 \leq l \leq q] \qquad (N \times q) \\
\boldsymbol{B} &= [b_{lk}, \quad 1 \leq l \leq q, \quad 1 \leq k \leq K] \qquad (q \times K) \\
\boldsymbol{\Phi} &= [\phi_k(t_j), \quad 1 \leq k \leq K, 1 \leq j \leq J] \qquad (N \times J) \\
\boldsymbol{\epsilon} &= [\epsilon_i(t_j), \quad 1 \leq i, \leq N, \quad 1 \leq j \leq J], \qquad (N \times J).
\end{aligned}$$

The penalized sum of squares is then defined as

$$\sum_{i=1}^{N}\sum_{j=1}^{J}(Y_i(t_j) - [\boldsymbol{X}\boldsymbol{B}\boldsymbol{\Phi}_{ij}])^2 + \sum_{l=1}^{q}\lambda_l \int \{(L\beta_l)(t)\}^2 \tag{6}$$

define the matrices $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_q)$ and $\boldsymbol{R} = [\int L(\phi_l)(t)L(\phi_k)(t)dt, \quad 1 \leq l, k \leq K]$. Also, let $vec(\boldsymbol{M})$ be a column vector of length pq constructed by stacking the columns of $\boldsymbol{M}$, placing the first column on top. Also let $\boldsymbol{A} \otimes \boldsymbol{B}$ be the Kronecker product between matrix $\boldsymbol{A}$ and matrix $\boldsymbol{B}$. The elements of the matrix $\boldsymbol{B}$ which minimizes Equation (6) can be expressed as

$$vec(\boldsymbol{B}^T) = [\boldsymbol{U}^T U + \boldsymbol{\Lambda} \otimes \boldsymbol{R}]^{-1}\boldsymbol{U}^T vec(\boldsymbol{Y}^T)$$

where $\boldsymbol{U} = \boldsymbol{X} \otimes \boldsymbol{\Phi}^T$. We will omit the proof of this, which is quite involved.

Exercise 26. *Compute the length of the vector $vec(\boldsymbol{B}^T)$ based on the above equation.* □

One common goal when we are using linear regression models is to test wether some regressors have effects on the data. Conisder the model in Equation 5, an consider, for $m < q$, that we wish to test

$$H_0 : \beta_{m+1}(t) = \cdots = \beta_q(t) = 0, \quad \forall t \in [0, 1]$$

First, consider the quantities

$$R_p(t) = \sum_{i=1}^{N}(Y_i(t) - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j^{(p)}(t))^2$$

and

$$R_m(t) = \sum_{i=1}^{N}(Y_i(t) - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j^{(m)}(t))^2$$

where $\hat{\beta}_j^{(p)}(t)$ and $\hat{\beta}_j^{(m)}(t)$ are the estimators of β_j using a model with p and m regressor respectively. Also, define

$$F(t) = \frac{R_m(t) - R_p(t)/(p-m)}{R_p(t)/(N-p)}$$

since the estimation of β_j generally is pretty involved and require penalties, $F(t)$ does not have an F distribution, even for a fixed t. But consider the statistic

$$\hat{F}_{sup} = \sup_{t\in[0,1]} \hat{F}(t)$$

The distribution of $\hat{F}_{sup}$ under H_0 is unknown but it can be approximated as follows. Consider the $N!$ permutations, π, of the indexes $\{1, 2, \ldots, N\}$ and consider the permutated dataset

$$D_\pi = (Y_{\pi(i)}(t), \quad t \in [0,1], \quad x_{\pi(i),1}, \ldots, x_{\pi(i),q})$$

So sing D_π, compute $\hat{\beta}^{(m)}_{\pi,j}(t)$ for $j = 1, \ldots, m$ and $\hat{\beta}^{(p)}_{\pi,j}(t)$ for $j = 1, \ldots, q$ and find the statistics $\hat{F}^{\pi}_{sup}$. The $N!$ values of $\hat{F}^{\pi}_{sup}$ are arranged from smallest to largest. Let $c_{0.95}(N)$ be the 95-th percentile of $\hat{F}^{\pi}_{sup}$. Then, the null hypothesis is rejected if $\hat{F}_{sup} > c_{0.95}(N)$. This is what we call in statistics a permutation test. You can check some standard textbook, like Manly (2018) for justification of these kinds of procedures.

The methods we have discussed in this subsection are implemented in the `refund` package.

Exercise 27. *This exercise illustrates other uses of permutation testing in functional data analysis Let's say we observe two samples of curves defined on the same interval, $X_1, \ldots X_N$ and $X_{N+1}, \ldots X_{N+M}$. where the first N curves are i.i.d draws from a population with mean function μ_1 and the last M curves are i.i.d draws from a population with mean function μ_2. We want to test*

$$H_0 : \mu_1(t) = \mu_2(t) \quad \forall t \in [0,1]$$

Define the test statistic

$$\hat{T}_{sup} = \sup_{t\in[0,1]} \hat{T}(t) = \frac{|\bar{X}_N(t) - \bar{X}_M(t)|}{\left\{N^{-1}\hat{V}_N(t) + M^{-1}\hat{V}_M(t)\right\}^{1/2}}$$

where $\bar{X}_N$ and $\bar{X}_M$ are the sample means for each set of curves, and $\hat{V}_N(t)$, $\hat{V}_M(t)$ are the sample variances. The null hypothesis is rejected if $\hat{T}_{sup}$ is large for some t. The distribution of $\hat{T}_{sup}$ is unknown but can be approximated via permutations. Let the indexes be $\{1, \ldots, N, N+1, \ldots, N+M\}$ and let π be a permutation on this indexes. If H_0 is true, the two samples $X_{\pi(1)}, \ldots, X_{\pi(N)}$ and $X_{\pi(N+1),\ldots,X_{\pi(N+M)}}$ have the same mean, and we can use $\hat{T}^{\pi}_{sup}$ as an approximation of the distribution of $\hat{T}_{sup}$. Apply this methodology then to test the null hypothesis

$$H_0 : \mu_M(T) = \mu_F(t)$$

where μ_M is the mean growth curve of males and μ_F is the mean growth curve of females in the Berkeley growth dataset curve. Plot $\hat{T}(t)$ and $c_{0.95}(N+M)$ on the same plot

□

6.4 The fully functional model

Now let's consider the fully functional linear model, or function-on-function regression model.

$$Y_i(t) = \alpha(t) + \int \psi(t,s)X_i(s)ds + \epsilon_i(t), \quad i = 1, \ldots N$$

The responses $Y_i(t)$ is a function, the regressor X_i is also a function. The regression coefficient is a bivariate function, or kernel, ψ. We can be a little more general and include some more functional regressors, as well as some scalar regressors. We will only consider two functional regressors, since more will burden our analysis with excessive notation, but keep in mind that the procedure is easily generalizable. Then, the model has the form

$$Y_i(t) = \alpha(t) + w_i^T\gamma + \int_{\mathcal{S}} \psi_1(t,s)X_{i1}(s)ds + \int_{\mathcal{R}} \psi_2(t,s)X_{i2}(r)dr + \epsilon_i(t) \tag{7}$$

so for each subject i, $1 \leq i \leq N$ we observe

$$Y_i(t_{ij}), \quad 1 \leq j \leq J_i; \qquad X_{i1}(s_{ik}), \quad 1 \leq k \leq K_j; \qquad X_{i2}(r_{il}), \quad 1 \leq l \leq L_i$$

as well as the values of the scalar regressors $\boldsymbol{w} = [w_{i1}, w_{i2}, \ldots, w_{iq}]^T$. The unknown parameters are $\gamma = [\gamma_1, \ldots, \gamma_q]^T$ and the bivariate functions ψ_1 and ψ_2. The response functions are defined on the compact interval $\mathcal{T}$, while the regressors on possibly different intervals $\mathcal{S}$ and $\mathcal{R}$. ϵ_i are assumed to be i.i.d. Gaussian random functions. Also, assume X_{i1} and X_{i2} are centered.

Also, assume that t_{ij}, s_{ik} and r_{il} are dense in their respective domains. If they are sparse, other methodologies should be used. Consider the simple setting in which the functions are observed on the same dense grids, $t_{ij} = t_j$, $s_{ik} = s_k$, $r_{il} = r_l$ for each subject. Now, let the intercept function be represented as

$$\alpha(t) = \sum_{m=1}^{M} \alpha_m B_m(t)$$

where $B_1, B_2, \ldots, B_M$ are some basis functions and M is sufficiently big. We also need to represent ψ_1, ψ_2 as a linear combination of basis. Note that they are Bivariate, so we need a bivariate basis system. These systems are generally constructed as products of univariate basis functions. We then use the approximations

$$\psi_1(t,s) = \sum_{g=1}^{G} \psi_{1,g} B_{1,g}(t,s)$$
$$\psi_2(t,r) = \sum_{h=1}^{H} \psi_{2,h} B_{2,h}(t,r)$$

Using this expansions, and approximating the integrals in Equation 7 as a Riemann sum,

$$\begin{aligned}
\int_{\mathcal{S}} \psi_1(t,s) x_{i1}(s) ds &= \sum_{k=1}^{K} (s_k - s_{k-1}) \psi_1(t, s_k) X_{i1}(s_k) \\
&= \sum_{k=1}^{K} (s_k - s_{k-1}) \sum_{g=1}^{G} \psi_{1,g} B_{1,g}(t, s_k) X_{i1}(s_k) \\
&= \sum_{g=1}^{G} B^*_{1,g,i}(t) \psi_{1,g}
\end{aligned}$$

where $B^*_{1,g,i} = \sum_{k=1}^{K} (s_k - s_{k-1}) B_{1,g}(t, s_k) X_{i1}(s_k)$. Analogously,

$$\int_{\mathcal{S}} \psi_2(t,r) x_{i2}(r) dr = \sum_{h=1}^{H} B^*_{2,h,i}(t) \psi_{2,h}$$

where $B^*_{2,h,i}(t) = \sum_{l=1}^{L} (r_l - r_{l-1}) B_{2,h}(t, r_l) x_{i2}(r_l)$. Note that $B^*_{1,g,i}$ and $B^*_{2,h,i}(t)$ are both known if the basis functions are chosen. The number of basis functions, G and H is typically large. Therefore, we can write Equation 7 as

$$Y_i(t_j) = w_i^T \gamma + \sum_{m=1}^{M} B_m(t_j) \alpha_m + \sum_{g=1}^{G} B^*_{1,g,i} \psi_{1,g} + \sum_{h=1}^{H} B^*_{2,h,i}(t) \psi_{2,h} + \epsilon_i(t)$$

End lecture six.

References

Akhiezer, N. and I. Glazman (2013). *Theory of Linear Operators in Hilbert Space*. Dover Books on Mathematics. Dover Publications.

Aykroyd, R. G., V. Leiva, and F. Ruggeri (2019). Recent developments of control charts, identification of big data sources and future trends of current research. *Technological Forecasting and Social Change 144*, 221–232.

Azcorra, A., L. F. Chiroque, R. Cuevas, A. F. Anta, H. Laniado, R. E. Lillo, J. Romo, and C. Sguera (2018, may). Unsupervised scalable statistical method for identifying infuential users in online social networks. *Scientific Reports 8*(6955), 7.

Bapna, R., W. Jank, and G. Shmueli (2008). Price formation and its dynamics in online auctions. *Decision Support Systems 44*(3), 641–656.

Buckner, R. L., D. Head, J. Parker, A. F. Fotenos, D. Marcus, J. C. Morris, and A. Z. Snyder (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. *NeuroImage 23*(2), 724–738.

Calhoun, V. and T. Adali (2012). Analysis of complex-valued functional magnetic resonance imaging data: are we just going through a" phase"? *Bulletin of the Polish Academy of Sciences. Technical Sciences 60*(3), 371–418.

Calle-Saldarriaga, A., H. Laniado, F. Zuluaga, and V. Leiva (2021). Homogeneity tests for functional data based on depth-depth plots with chemical applications. *Chemometrics and Intelligent Laboratory Systems 219*, 104420.

Chapados, C. and D. J. Levitin (2008). Cross-modal interactions in the experience of musical performances: Physiological correlates. *Cognition 108*(3), 639–651.

Claeskens, G., M. Hubert, L. Slaets, and K. Vakili (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association 109*(505), 411–423.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference 147*, 1–23.

Cuevas, A., M. Febrero, and R. Fraiman (2004). An anova test for functional data. *Computational Statistics & Data Analysis 47*(1), 111–122.

d. Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer Verlag.

Dabo-Niang, S., F. Ferraty, and P. Vieu (2007). On the using of modal curves for radar waveforms classification. *Computational Statistics & Data Analysis 51*(10), 4878–4890.

Debnath, L. and P. Mikusinski (2005). *Introduction to Hilbert Spaces with Applications*. Elsevier Science.

Delicado, P., R. Giraldo, C. Comas, and J. Mateu (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics 21*(3-4), 224–239.

Durá Gil, J., J. M. Belda Lois, R. Poveda, A. Page, J. Laparra Hernández, J. Das, J. Prat, and A. García Belenguer (2010, 05). Comparison of functional regression and nonfunctional regression approaches to the study of the walking velocity effect in force platform measures. *Journal of applied biomechanics 26*, 234–9.

Erbas, B., M. Akram, D. Gertig, D. English, J. Hopper, A. Kavanagh, and R. Hyndman (2010, 02). Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the united states and england–wales. *Journal of epidemiology / Japan Epidemiological Association 20*, 159–65.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer New York.

Fraiman, R. and G. Muniz (2001). Trimmed means for functional data. *TEST 10*(2), 419–440.

Goldsmith, J. and F. Scheipl (2014). Estimator selection and combination in scalar-on-function regression. *Computational Statistics & Data Analysis 70*, 362–372.

Górecki, T., M. Krzyśko, Ł. Waszak, and W. Wołyński (2018). Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers 59*(1), 153–182.

Grenander, U. (1950, October). Stochastic processes and statistical inference. *Arkiv for Matematik 1*(3), 195–277.

Henderson, B. (2006). Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics 17*(1), 65–80.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer New York.

Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley.

Hutchinson, R., P. McLellan, J. Ramsay, H. Sulieman, and D. Bacon (2004). Investigating the impact of operating parameters on molecular weight distributions using functional regression. *Macromolecular Symposia 206*(1), 495–508.

Hyndman, R. J. and H. L. Shang (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics 19*(1), 29–45.

James, G. M. and T. J. Hastie (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 63*(3), 533–550.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. Chapman & Hall / CRC numerical analysis and scientific computing. CRC Press.

Kowal, D. R. (2019). Integer-valued functional data analysis for measles forecasting. *Biometrics 75*(4), 1321–1333.

Lee, K. L., R. J. Meyer, and E. T. Bradlow (2009). Analyzing risk response dynamics on the web: The case of hurricane katrina. *Risk Analysis 29*(12), 1779–1792.

Li, P.-L. and J.-M. Chiou (2011, jun). Identifying cluster number for subspace projected functional data clustering. *Comput. Stat. Data Anal. 55*(6), 2090–2103.

Lucero, J. C. (1999). Computation of the harmonics-to-noise ratio of a voice signal using a functional data analysis algorithm. *Journal of Sound and Vibration 222*, 512–520.

Manly, B. F. (2018). *Randomization, Bootstrap and Monte Carlo Methods in Biology: Texts in Statistical Science*. chapman and hall/CRC.

Marron, J. S., J. O. Ramsay, L. M. Sangalli, and A. Srivastava (2015). Functional Data Analysis of Amplitude and Phase Variation. *Statistical Science 30*(4), 468 – 484.

Maslova, I., P. Kokoszka, J. Sojka, and L. Zhu (2010). Statistical significance testing for the association of magnetometer records at high-, mid- and low latitudes during substorm days. *Planetary and Space Science 58*(3), 437–445.

McLean, M. W., G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics 23*(1), 249–269. PMID: 24729671.

Müller, H.-G. (2016). Peter Hall, functional data analysis and random objects. *The Annals of Statistics 44*(5), 1867 – 1887.

Müller, H.-G., S. Wu, A. D. Diamantidis, N. T. Papadopoulos, and J. R. Carey (2009). Reproduction is adapted to survival characteristics across geographically isolated medfly populations. *Proceedings of the Royal Society B: Biological Sciences 276*(1677), 4409–4416.

Newell, J., K. McMillan, S. Grant, and G. McCabe (2006). Using functional data analysis to summarise and interpret lactate curves. *Computers in Biology and Medicine 36*(3), 262–275.

Ogden, R. T., C. E. Miller, K. Takezawa, and S. Ninomiya (2002). Functional regression in crop lodging assessment with digital images. *Journal of Agricultural, Biological, and Environmental Statistics 7*(3), 389–402.

Panaretos, V. M. and S. Tavakoli (2013). Fourier analysis of stationary time series in function space. *The Annals of Statistics 41*(2), 568–603.

Patil, A. A., J. Bovy, G. Eadie, and S. Jaimungal (2022, feb). Functional data analysis for extracting the intrinsic dimensionality of spectra: Application to chemical homogeneity in the open cluster m67. *The Astrophysical Journal 926*(1), 51.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. Use R! Springer New York.

Ramsay, J., J. Ramsay, B. Silverman, S. S. Media, and H. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Ramsay, J. O. (1982). When the data are functions. *Psychometrika 47*(4), 379–396.

Ramsay, J. O., S. Graves, and G. Hooker (2021). *fda: Functional Data Analysis*. CRAN. R package version 5.4.0.

Ramsay, J. O., G. Hooker, D. Campbell, and J. Cao (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(5), 741–796.

Ramsay, J. O. and J. B. Ramsey (2002). Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics 107*(1), 327–344. Information and Entropy Econometrics.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics 14*(1), 1–17.

Reiss, P. T., J. Goldsmith, H. L. Shang, and R. T. Ogden (2017). Methods for scalar-on-function regression. *International Statistical Review 85*(2), 228–249.

Robinson, J. and A. H. Welsh (2018). Peter gavin hall. 20 november 1951—9 january 2016. *Biographical Memoirs of Fellows of the Royal Society 64*, 207–229.

Rupp, A. A. (2005). Quantifying subpopulation differences for a lack of invariance using complex examinee profiles: An exploratory multigroup approach using functional data analysis. *Educational Research and Evaluation 11*(1), 71–97.

Shang, H. L. and R. J. Hyndman (2018). *fds: Functional Data Sets*. Macquarie University. R package version 1.8.

Staicu, A.-M., C. M. Crainiceanu, D. S. Reich, and D. Ruppert (2012). Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics 68*(2), 331–343.

Stewart, K. J., D. P. Darcy, and S. L. Daniel (2006). Opportunities and challenges applying functional data analysis to the study of open source software evolution. *Statistical Science 21*(2), 167–178.

Torres, J. M., P. G. Nieto, L. Alejano, and A. Reyes (2011). Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of Hazardous Materials 186*(1), 144–149.

Ullah, S. and C. F. Finch (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology 13*(43), 1–12.

Vidal, M., M. Rosso, and A. M. Aguilera (2021). Bi-smoothed functional independent component analysis for eeg artifact removal. *Mathematics 9*(11), 1243.

Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Functional data analysis. *Annual Review of Statistics and Its Application 3*(1), 257–295.

Wu, P.-S. and H.-G. Müller (2010, 01). Functional embedding for the classification of gene expression profiles. *Bioinformatics 26*(4), 509–517.

Yao, F., H.-G. Müller, and J.-L. Wang (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*(470), 577–590.

Yao, F., H.-G. Müller, and J.-L. Wang (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics 33*(6), 2873 – 2903.

Zhang, C., Y. Chai, X. Guo, M. Gao, D. Devilbiss, and Z. Zhang (2016). Statistical learning of neuronal functional connectivity. *Technometrics 58*(3), 350–359.

Zhang, Y., H.-G. Müller, J. R. Carey, and N. T. Papadopoulos (2006). Behavioral trajectories as predictors in event history analysis: Male calling behavior forecasts medfly longevity. *Mechanisms of Ageing and Development 127*(8), 680–686.

Zhang, Z. and B. Saparbayeva (2022). Amplitude mean of functional data on s2 and its accurate computation. *Journal of Mathematical Imaging and Vision 1*(1), 1–19.