PLACEHOLDER-TITLE: Functional Linear Regression in a Scalar-on-Function Setting with Applications to SOMETHING

Jonghun Baek, Jakob Juergens, Jonathan Willnow

whenever

Research Module in Econometrics and Statistics
Winter Semester 2021/2022

Contents

1 Colour Guide

- RED: is for general comments for your own text
- GREEN: is for Jona's comments
- ORANGE: is for Jonghun's comments
- BLUE: is for Jakob's comments

2 Introduction

- Describe the idea of regressing a scalar on functional data
- Describing the difference to multiple linear regression intuitively
- Giving an intuitive example

Functional Data Analysis (FDA) is a relatively new field (roots in the 1940s Grenander and Karhunen) which is gaining more attention as researchers from different fields collect data that is functional in nature. Classical statistical methods can often process this data, but only FDA allows extracting information given by the smoothness of the underlying process (cf. Levitin et al. 2007). As Kokoszka and Reimherr 2017 describe, FDA should be considered when one can view variables or units of a given data set as smooth curves or functions and the interest is in analyzing samples of curves (cf. Kokoszka and Reimherr 2017, S. 17).
To motivate scalar-on-function regression, consider the case of a data set containing a scalar response and observations of an underlying continuous process. In economics, one application could be the regression of stock market correlations on the Global Crisis Index (GCI), where the regression allows to assess the relationship between the correlation and the GCI at every point within a window (cf. Das et al. 2019).
The focus of this paper is to introduce Functional Linear Regression (FLR) in a scalar-on-function setting. We will be using the standard FLR framework, which relates functional predictors to a scalar response as follows: (I don't set up any interval for s here we might do later...)

$$Y_i = \beta_0 + \int X_i(s)\beta(s)ds + \epsilon_i, \qquad i = 1, ..., n \tag{1}$$

where the X_i are realizations of a random function $\mathbf{X}$, Y_i are the corresponding realizations of the response variable and $\beta(s)$ is the coefficient function. The distinct feature of this framework is that the regressor is a function, which necessitates a different approach to estimation. As in the well-known framework of scalar linear regression, this is motivated by an interest in $\beta(s)$ for prediction. For instance, fluctuation in $X_i(s)$ at a point s_0 will not have any effect on Y_i if $\beta(s_0) = 0$.
Estimation of $\beta(s)$ is inherently an infinite-dimensional problem. In Section 2, after introducing the necessary theoretical concepts, we describe three methods of estimating a scalar coefficient function using a concept called truncated basis expansion. We report the results of the Monte-Carlo simulation regarding these three different methods in Section 3. Finally, in Section 4, we test the prediction of FLR in a real-world setting. (We may put some simple descriptions of results about each of MC and Application)

3 Theory

In multivariate regression, data is often observed in the form of elements from Euclidean space, $\mathbb{R}^p$. However, the statistics derived from infinite-dimensional random functions cannot be defined on a finite dimensional space. To understand functional linear regression and the differences between the methods presented in this paper, it is therefore necessary to introduce some concepts and extend known aspects of linear regression theory to include functional objects. One integral concept in inferential statistics are random variables. Paraphrasing a definition by Bauer 2020, a random variable $X : \Omega \to \Omega'$ is an $\mathcal{A}$-$\mathcal{A}'$-measurable function, where $(\Omega, \mathcal{A}, P)$ is a probability space and $(\Omega', \mathcal{A}')$ is a measure space.
A typical case known to every undergraduate student of economics in less formal detail is $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the canonical σ-algebra on the real numbers. As a first intuition, it is possible to imagine a similar concept where a random variable does not realize as an element of the real numbers but as a function in a function space. A formalization of this idea makes some more theoretical considerations necessary. The following theoretical introduction closely follows chapters 2.3 and 2.4 from Hsing and Eubank 2015.

3.1 Inner Products and Hilbert Spaces

Let $\mathbb{V}$ be a vector space over some field of scalars $\mathbb{F}$. A function $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{F}$ is called an inner product, if $\forall v, v_1, v_2 \in \mathbb{V}$ and $a_1, a_2 \in \mathbb{F}$ the following properties hold.

1. $\langle v, v \rangle \geq 0$
2. $\langle v, v \rangle = 0$ if $v = 0$
3. $\langle a_1 v_1 = a_2 v_2, v \rangle = a_1 \langle v_1, v \rangle + a_2 \langle v_2, v \rangle$
4. $\langle v_1, v_2 \rangle = \langle v_2, v_1 \rangle$

A vector space with an associated inner product is called an inner product space. [verbatim quote!] The inner product naturally defines a norm and an associated distance on the vector space as follows. In the following we restrict our analysis to the case of $\mathbb{F} = \mathbb{R}$.

$$||v|| = \langle v, v \rangle^{\frac{1}{2}} \tag{2}$$

$$d(v_1, v_2) = \langle v_2 - v_1, v_2 - v_1 \rangle^{\frac{1}{2}} \tag{3}$$

If the inner product space is complete with respect to the induced distance, it is called a Hilbert space, denoted $\mathbb{H}$ in the following. To extend the known concept of a basis in a finite dimensional space to the potentially infinite Hilbert spaces, it is necessary to define the closed span of a sequence of elements of $\mathbb{H}$. Recall that the span of a set of vectors $S \subseteq \mathbb{R}^P$ is given by

$$span(S) = \left\{ \sum_{i=1}^{k} \lambda_i v_i \;\middle|\; k \in \mathbb{N},\ v_i \in S,\ \lambda_i \in \mathbb{R} \right\} \tag{4}$$

The closed span $\overline{span}(S)$ of a sequence S in $\mathbb{H}$ is defined as the closure of the span with respect to the distance induced by the norm. S is called a basis of $\mathbb{H}$ if $\overline{span}(S) = \mathbb{H}$.
It is called an orthonormal basis, if in addition the following properties hold.

1. $\langle v_i, v_j \rangle = 0 \quad \forall v_i, v_j \in S \quad i \neq j$
2. $||v|| = 1 \quad \forall v \in S$

3.2 Hilbert Space of Square-Integrable Functions

In functional data analysis, one Hilbert space of particular importance is the space of square-integrable functions on $[0,1]$ denoted $\mathbb{L}^2[0,1]$. To define it, look first at the measure space given by $([0,1], \mathcal{B}, \mu)$ where $\mathcal{B}$ is the Borel σ-algebra on $[0,1]$ and μ is the Lebesgue-measure.
Then $\mathbb{L}^2[0,1]$ is the collection of all measurable functions f on $[0,1]$ that fulfill the following condition.

$$\int_0^1 |f|^2 d\mu < \infty \tag{5}$$

Its inner product is defined as

$$\langle f_1, f_2 \rangle = \int_0^1 f_1 f_2 d\mu. \tag{6}$$

The Hilbert space of square integrable functions on $[0,1]$ is the function space that is most often used for theoretical considerations in functional data analysis. Trivially, it is possible to extend this to other closed intervals on the reals line, but there are also more complex generalizations. For the purpose of this paper we will focus on the typical case and assume that random functions are random variables realizing in $\mathbb{L}^2$ for some closed interval of the real numbers.

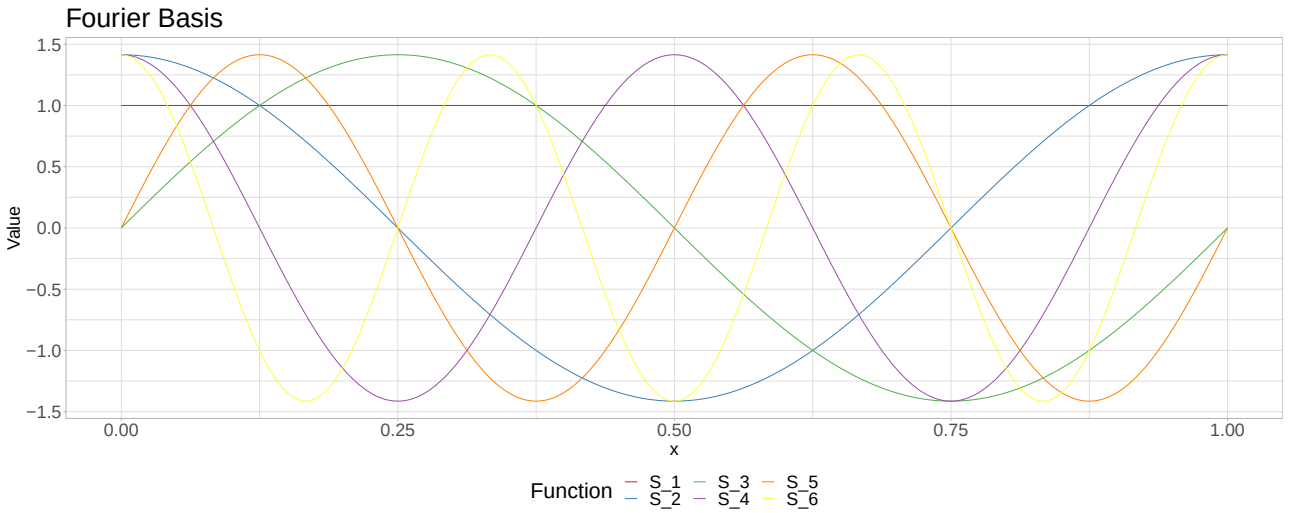
3.3 Different Bases of $\mathbb{L}^2$

Two examples of orthonormal bases of $\mathbb{L}^2[0,1]$ that are often used in practice are explained in the following.

b-spline Basis

Fourier Basis The Fourier basis for $\mathbb{L}^2[0,1]$ is given by the following sequence of functions.

$$S_{FB,i}(x) = \begin{cases} 1 & \text{if} \quad i = 1 \\ \sqrt{2}\cos(\pi i x) & \text{if} \quad i \quad \text{is even} \\ \sqrt{2}\sin(\pi(i-1)x) & \text{otherwise} \end{cases} \tag{7}$$



3.4 Karhunen-Loéve Expansion and Empirical Eigenbases

3.5 Detailed Draft

- Motivate random functions from introduction and the general concept of random variables
- Formalize random function in this context as random variables realizing in a Hilbert space
- Introduce $\mathbb{L}^2[0,1]$ as the Hilbert space of square integrable functions on $[0,1]$
- Specialize to Hilbert space being $\mathbb{L}^2[0,1]$ for this context
- Define mean and covariance function of a random function realizing in $\mathbb{L}^2[0,1]$
- Introduce the concept of a basis of a Hilbert space and specialize to $\mathbb{L}^2[0,1]$
- Introduce b-spline and Fourier bases
- Introduce eigenfunctions and FPCA on the basis of covariance function (Karhunen-Loéve expansion)
- explain similarities to Eigenvalues and Eigenvectors of matrix + PCA (fraction of explained variance etc...)
- Introduce functional observations in this context as realizations of a random variable realizing in $\mathbb{L}^2[0,1]$
- Explain the concept of iid data in a functional setting

- Define point-wise mean (sample), point-wise standard deviation (sample) and sample covariance function
- Explain approximations of functional observations using truncated basis representations
- Introduce linear operator L_1 and sufficient condition associated with it
- Motivate Scalar-on-function regression from multivariate linear regression with a scalar response variable

There are several important aspects of functional regression in this functional setting that separate it from usual multiple regression according to Kokoszka and Reimherr 2017: In functional linear regression, the aim is not only to obtain an estimate of the function $\beta(s)$ — this estimate also needs to have a useful interpretation. Without it, there might be prediction, but the increase in understanding of the underlying question will be minimal. One aspect of a useful interpretation is that the estimate $\beta(s)$ should not jump in a seemingly random fashion, because an interpretation of this erratic behavior will often be impossible.

- Explain problem of naively extending multivariate linear regression to infinite dimensions

A common setting in non-functional regression is akin to the following. Assume a model as follows:

$$Y = X'\beta + \epsilon \tag{8}$$

where $X \in \mathbb{R}^{n \times J}$ is a matrix containing the regressors, $\beta \in \mathbb{R}^J$ is a coefficient vector and *epsilon* is a vector containing the error term. For simplicity, assume that the data generating process fulfills the Markov assumptions. Then the famous OLS-estimator is given by:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \tag{9}$$

A naive approach to FLR would be to try to generalize this to the functional setting. Assuming a data generating process of the form:

$$Y = \int \beta(s)X(s)\,dt \; + \epsilon \tag{10}$$

it becomes clear that we cannot compute the estimate of $\beta(t)$ as we would do in a classical multivariate setup because of the infinite dimensionality of the underlying objects. Where in the finite dimensional setting the OLS estimator can be derived as a method of moments estimator by solving a system of equations of sample moment restrictions, this leads to a system of infinitely many equations in the functional setting. In practice, functional observations are never truly continuously observed. If we assume that the functional observations are observed at a finite set of points $\{t_1, \ldots, t_J\}$ this makes the derivation of an OLS estimator possible as before.

$$Y_i = \sum_{j=1}^{n} \beta(t_j)X_i(t_j) + \epsilon_i, \tag{11}$$

However, this often still results in a large and difficult to solve system of equations. Even if solved, the result is often a noisy function $\hat{\beta}(s)$ that is not useful for interpretation since it does not use the intuition of smooth functions. Another reason why estimation is not feasible using this approach is colinearity. Looking at equation 11 and assuming continuous functions X_i it becomes clear that if t_j is close to $t_{j'}$, $X_i(t_j)$ is close to $X_i(t_{j'})$. Thereby, there will be vectors $X_i = (X_i(t_1), \ldots, X_i(t_J))'$ that are highly correlated and thus lead to large variances of β. (cf. Kokoszka and Reimherr 2017)
A different approach is necessary.
Define therefore

$$c_{\mathbf{X}}(t,s) = E[\mathbf{X}(t)\mathbf{X}(s)],\; c_{\mathbf{XY}}(t) = E[\mathbf{X}(t)\mathbf{Y}], \tag{12}$$

Under the assumption that X is independent from ϵ we obtain

$$c_{\mathbf{XY}}(t) = E[\mathbf{X}(t) \int \beta(t)\mathbf{X}(s)\,ds \; + \epsilon] \tag{13}$$

$$c_{\mathbf{XY}}(t) = E[\int \beta(s)\; \mathbf{X}(s)\mathbf{X}(t)\,ds \mid X] + E[\epsilon|\mathbf{X}] \tag{14}$$

$$c_{\mathbf{XY}}(t) = \int c_{\mathbf{X}}(t,s)\beta(s)\,ds \tag{15}$$

In practice, this results in a large and often difficult to solve system of equations. If solved, a perfect fit is possible,but results in a noisy and erratic function $\hat{\beta}(s)$ that is not useful for interpretation since it does not utilize the intuition of smooth functions (cf. Horváth and Kokoszka 2012). Another reason why

estimation is not feasible using this approach is colinearity. If we approximate the scalar-on-functional regression by assuming a set of discrete observation points for all realizations of the data generating process as

$$Y_i = \sum_{j=1}^{n} \beta(t_j) X_i(t_j) + \epsilon_i, \tag{16}$$

it becomes clear that if t_j is close to $t_{j'}$, $X_i(t_j)$ is close to $X_i(t_{j'})$ there will be vectors $X_i = (X_i(t_1), \ldots, X_i(t_J))'$ that are highly correlated and thus lead to large variances of β. (cf. Kokoszka and Reimherr 2017) isn't something missing here? Like "and employ standard multivariate linear regression" // Will be done (Jona)

- Solution: estimation using truncated basis expansion to approximate data (theoretical description)

The simplest approach to regularize the noisy estimate of $\hat{\beta}(s)$ is to expand it with deterministic basis function. Assume

$$\beta(t) = \sum_{k=1}^{K} c_k B_k(t) \tag{17}$$

to expand

$$\int \beta(s) X(s)\, dt \; = \sum_{k=1}^{K} c_k \int B_k(t) X_i(t)\, dt =: \sum_{k=1}^{K} x_{ik} c_k \tag{18}$$

to the linear model of equation 2 with $\mathbf{c} = [\alpha, c_1, c_2, ..., c_K]^T$ (corresponding to β) estimated by $\hat{\mathbf{c}} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y}$. Hence, the estimate $\hat{\beta}$ depends on the basis function B_k and its corresponding shape, where K is a subjective choice of the researcher. This subjective choice highly depends on the researchers intuition of the smoothness of the estimate. It is best practice to use the same basis functions as in add link to smoothing (cf. Kokoszka and Reimherr 2017)
include Part about CI´s? its about inference and not about prediction

- Problem: truncation error δ and how to deal with it?
- Explain how to address truncation error in standard errors
- Motivate three estimation procedures
    1. truncated b-spline basis expansion without addressing truncation error
    2. truncated b-spline basis expansion WITH addressing truncation error
    3. truncated Eigenbasis expansion (advantages: low number of basis functions get low approximation error)

3.6 Draft-Overview

- Motivate Karhunen-Loeve-Expansion and Eigenbasis from PCA
- Explain Scalar-on-Function Regression
- Estimation through basis-expansion (incl. Eigenbasis) [and estimation with roughness penalty]
- Address approximation error due to basis-truncation

3.7 Literature

- Kokoszka and Reimherr 2017
- Hsing and Eubank 2015
- Ramsay and Silverman 2005
- Horváth and Kokoszka 2012
- Cai and Hall 2006
- Levitin et al. 2007

4 Simulation Study

4.1 Draft-Overview

- Motivate Simulation for some data generating process from application
- Describe Simulation Setting from technical standpoint (DGP, set-up for replication, ...)
- Compare estimation with
    1. b-spline basis without addressing approximation error
    2. ... including proper treatment of approximation error
    3. Eigenbasis constructed from observations
- Prediction not Inference (Alternative: Focused on a testing procedure motivated by the application)
- Present Results
- Explain relevance for application

4.2 Motivate Simulation for some data generating process from application

For the simulation study, we use the gasoline dataset to predict the octane ratings of gasoline samples relying on the introduced methods. This has become more relevant as modern internal combustion engines become more complex and rely on precisely tuned fuels. The dataset, which contains 60 i.i.d. observations with each 400 measurements, was constructed using Near infra-red (NIR) spectroscopy, which allows to analyse samples of gasoline much faster and with the same reproducibility as standard tests (cf. Gy. Bohács, Z. Ovádi, A. Salgó 1998). The study follows Reiss and Ogden (2007) as a guideline. Similar to Reiss and Ogden (2007), two different true coefficient functions f_1 and f_2 were chosen that differ in their smoothness:

$$f_1 = 2\sin(0.5\pi t) + 4\sin(1.5\pi t) + 5\sin(2.5\pi t) \tag{19}$$

$$f_2 = 1.5^{\frac{-0,5(t-0.3)^2}{0.02^2}} - 4^{\frac{-0,5(t-0.45)^2}{0.015^2}} + 8^{\frac{-0,5(t-0.6)^2}{0.02^2}} - 1^{\frac{-0,5(t-0.8)^2}{0.03^2}} \tag{20}$$

Two different error-terms ϵ were created by first generating iid standard normal error and then multiplying them by σ_e which is calculated such that the squared multiple correlation coefficient $R^2 = var(Xf)/(var(Xf) + \sigma_e^2)$ is equal to 0.9 and 0.6. The two error-terms are then computed to generate two sets of responses with different signal-to-noise ratios for each true function, using the gasoline dataset. These four combinations are then used with different number of basis-function $(5, 6, ..., 25)$ of the order 5 to predict the responses using the b-spline basis approach and the FPCR approach. Within one repetition, the data is randomly sampled into a training and a test set to calculate the reported test MSE. The simulation was done with R (version...). In total we carried out 2000 simulations for each combination of data and number of basis functions. Information about version and link to repo / package in footnote??

4.3 Results

4.3.1 Interpretation and Relevance for Application

4.4 Literature

- Shonkwiler and Mendivil 2009
- R-packages: fda, refund, mgcv

5 Application

The application uses the insights from the previous sections to predict the octane ratings of the introduced gasoline dataset. Following the results from the simulation study,...

- incorporate results of simulation study: number of basis, components, etc...
- point out difficulties estimating sderror
- describe setup and results

5.1 Draft-Overview

- Prediction not Inference (Alternative: Focused on a testing procedure motivated by the data set)
- IID data set (no dependence between the curves, don't want to do functional time series)
- Not necessarily data from economics (like biology, sports, whatever)
- Smooth curves or random walk (both fine)
- https://functionaldata.wordpress.ncsu.edu/resources/

5.2 Literature

- Carey et al. 2002

6 Outlook

6.1 Literature

- James, Wang, and Zhu 2009 (shape-restrictions)

7 Appendix

8 Bibliography

Bauer, Heinz (May 2020). *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie*. de. Publication Title: Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie. De Gruyter. ISBN: 978-3-11-231316-9. DOI: 10.1515/9783112313169. URL: https://www.degruyter.com/document/doi/10.1515/9783112313169/html (visited on 11/13/2021).

Cai, T. Tony and Peter Hall (Oct. 2006). "Prediction in functional linear regression". In: *The Annals of Statistics* 34.5. Publisher: Institute of Mathematical Statistics, pp. 2159–2179. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053606000000830. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-34/issue-5/Prediction-in-functional-linear-regression/10.1214/009053606000000830.full (visited on 10/24/2021).

Carey, James R. et al. (2002). "Life history response of Mediterranean fruit flies to dietary restriction". en. In: *Aging Cell* 1.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1474-9728.2002.00019.x, pp. 140–148. ISSN: 1474-9726. DOI: 10.1046/j.1474-9728.2002.00019.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1474-9728.2002.00019.x (visited on 10/24/2021).

Das, Sonali et al. (2019). "The effect of global crises on stock market correlations: Evidence from scalar regressions via functional data analysis". In: *Structural Change and Economic Dynamics* 50, pp. 132–147. ISSN: 0954-349X. DOI: 10.1016/j.strueco.2019.05.007. URL: https://www.sciencedirect.com/science/article/pii/s0954349x19301407.

Gy. Bohács, Z. Ovádi, A. Salgó (1998). "Prediction of Gasoline Properties with near Infrared Spectroscopy". In: *Journal of near infrared spectroscopy.* 6, pp. 341–348.

Horváth, Lajos and Piotr Kokoszka (May 2012). *Inference for Functional Data with Applications*. en. Google-Books-ID: OVezLB_-ZpYC. Springer Science & Business Media. ISBN: 978-1-4614-3655-3.

Hsing, Tailen and Randall Eubank (Mar. 2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. en. Google-Books-ID: om9uBwAAQBAJ. John Wiley & Sons. ISBN: 978-1-118-76256-1.

James, Gareth M., Jing Wang, and Ji Zhu (2009). "Functional linear regression that's interpretable". In: *The Annals of Statistics* 37.5A. ISSN: 0090-5364. DOI: 10.1214/08-AOS641.

Kokoszka, Piotr and Matthew Reimherr (Aug. 2017). *Introduction to Functional Data Analysis*. Englisch. 1st ed. Boca Raton: Chapman and Hall/CRC. ISBN: 978-1-4987-4634-2.

Levitin, Daniel et al. (Aug. 2007). "Introduction to Functional Data Analysis". In: *Canadian Psychology/Psychologie canadienne* 48, pp. 135–155. DOI: 10.1037/cp2007014.

Ramsay, James and B. W. Silverman (2005). *Functional Data Analysis*. en. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag. ISBN: 978-0-387-40080-8. DOI: 10.1007/b98888. URL: https://www.springer.com/de/book/9780387400808 (visited on 10/23/2021).

Shonkwiler, Ronald W. and Franklin Mendivil (2009). *Explorations in Monte Carlo Methods*. en. Undergraduate Texts in Mathematics. New York: Springer-Verlag. ISBN: 978-0-387-87836-2. DOI: 10.1007/978-0-387-87837-9. URL: https://www.springer.com/gp/book/9780387878362 (visited on 10/23/2021).