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 Introduction **3**

2 Theory **3**

2.1 Detailed Draft 3

2.2 Kokoszka Reimherr (2017) p51-53 4

2.3 Draft-Overview 5

2.4 Literature 5

3 Simulation **7**

3.1 Draft-Overview 7

3.2 Literature 7

4 Application **8**

4.1 Draft-Overview 8

4.2 Literature 8

5 Outlook **8**

5.1 Literature 8

6 Appendix **8**

7 Bibliography **9**

1 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 getting more attention as researchers from different fields collect more data from a continuous underlying process. This data still can be processed by classical statistical methods, but only FDA allows answering questions that are tied to the information given by the smoothness of the underlying continuous process (cf. Levitin et al. 2007).

As Kokoszka and Reimherr 2017 describe, FDA should be considered when one can view one or more of the variables or units of a given data set as a smooth curve or function 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 a continuous underlying 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 terms of scalar-on-function. We will be importing the standard FLR model, which shows functional predictors to a scalar response as follows: (I don't set up any interval for s here we might do later...)

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

where $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ is the realization of data, where X_i's are independent and identically distributed random function X and $\beta(s)$ is the coefficient function. The distinct is that the estimator is not a scalar but a function, which leads to compelling interest in $\beta(s)$ for prediction. The information about the function comes up with how large or small a future observation x of X will influence the response with the leverage on $\int \beta(s)xds$. For instance, fluctuation in X does not have any effect on Y, where $\beta(s) = 0$, or has a greater effect on Y with larger $\beta(s)$. Additionally, assume that $\beta(s)$ is exactly a linear function for any interval. The effect of X is, then, constant on that interval.

Estimation of $\beta(s)$ is inherent in a problem of infinite dimension. In Section 2, after constructing necessary theoretical properties to understand FDA, we progress to reduce dimension by utilizing smoothing with two types of basis function, namely, b-spline and eigenfunction. The results of the Monte-Carlo simulation regarding three different situations are reported in Section 3. Finally, in Section 4, we test the prediction of FLR with the actual data set. (We may put some simple descriptions of results about each of MC and Application)

2 Theory

2.1 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 $\mathbf{L}^2[0,1]$ as the Hilbert space of square integrable functions on $[0,1]$
- Specialize to Hilbert space being $\mathbf{L}^2[0,1]$ for this context
- Define mean and covariance function of a random function realizing in $\mathbf{L}^2[0,1]$
- Introduce the concept of a basis of a Hilbert space and specialize to $\mathbf{L}^2[0,1]$
- 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 $\mathbf{L}^2[0,1]$
- 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

2.2 Kokoszka Reimherr (2017) p51-53

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 regression, the aim is not only to compute an estimate of the function β because this function needs also to have an useful interpretation. Without this useful interpretation, there can be no effective and feasible prediction of the scalar responses from new explanatory functions. Hereby applies, that intervals with larger values of $|\beta(s)|$ are contributing more to to the response than small values of $|\beta(s)|$. The sign of $\beta(s)$ within the intervals of the value s show either negative or positive association of $|\beta(s)|$ for this interval. To get an useful interpretation, the estimate β cannot jump in a seemingly random fashion, because then an useful interpretation is not possible and predictions from this model tend to have large variances and center around the mean of the responses. (see Kokoszka and Reimherr 2017)

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

Considering the population model of scalar-on-function linear regression

$$Y = \int \beta(s)X(s)\,dt \; + \epsilon \tag{1}$$

it becomes clear that we cannot compute the estimate of $\beta(t)$ as we would do in a classical multivariate setup because there are infinitely many solutions for finding the minimizing argument for $\hat{\beta}$. Define

$$c_X(t,s) = E[X(t)X(s)],\ c_{XY}(t) = E[X(t)Y]. \tag{2}$$

Under the assumption that X is independent from ϵ we obtain

$$c_{XY}(t) = E[X(t)\int \beta(s)X(s)\,ds\ \ + \epsilon] \tag{3}$$

$$c_{XY}(t) = E[\int \beta(s)X(s)X(t)\,ds \mid X] + E[\epsilon|X] \tag{4}$$

$$c_{XY}(t) = \int c_X(t,s)\beta(s)\,ds \tag{5}$$

This results in practice in a large number of equations, which are difficult to solve. Even if solved, this results in a noisy function $\beta(\dot{})$ that is not useful for interpretation since it does bot utilize the intuition of smooth functions. Another reason why estimation is not feasible using this approach is colinearity. Approximate the scalar-on-functional regression as

$$Y_i = \sum_{i=1}^{n} \beta(t_i)X_1(t_i) + \epsilon_i. \tag{6}$$

It becomes obvious that if t_i is close to $t_{i'}$, $X_i(t_i)$ is close to $X_i(t_{i'})$, so there will be vectors X_i thatare stronlgx correlated and thus lead to large variances and not feasible estimation.

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

2.3 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

2.4 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

3 Simulation

3.1 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

3.2 Literature

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

4 Application

4.1 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/

4.2 Literature

- Carey et al. 2002

5 Outlook

5.1 Literature

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

6 Appendix

7 Bibliography

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.

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).