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

- 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. This data can often be processed by classical statistical methods, but only FDA allows extracting information given by the smoothness of the underlying process (cf. **levitin˙introduction˙2007**). As **kokoszka˙introduction˙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˙introduction˙2017**).
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˙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. The results of the Monte-Carlo simulation regarding these three different methods are reported 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)

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 $\mathbb{L}^2[0,1]$ 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

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˙introduction˙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:

$$Y = X'\beta + \epsilon \tag{2}$$

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:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \tag{3}$$

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

$$Y = \int \beta(s)X(s)\,dt \; + \epsilon \tag{4}$$

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.

$$Y_i = \sum_{j=1}^{n} \beta(t_j)X_i(t_j) + \epsilon_i, \tag{5}$$

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 **??** 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˙introduction˙2017**)
A different approach is necessary.

Define therefore

$$c_{\mathbf{X}}(t,s) = E[\mathbf{X}(t)\mathbf{X}(s)],\; c_{\mathbf{XY}}(t) = E[\mathbf{X}(t)\mathbf{Y}], \tag{6}$$

Under the assumption that X is independent from ϵ we obtain

$$c_{\mathbf{XY}}(t) = E[\mathbf{X}(t) \int \beta(t)\mathbf{X}(s)\,ds \; + \epsilon] \tag{7}$$

$$c_{\mathbf{XY}}(t) = E[\int \beta(s)\; \mathbf{X}(s)\mathbf{X}(t)\,ds \mid X] + E[\epsilon|\mathbf{X}] \tag{8}$$

$$c_{\mathbf{XY}}(t) = \int c_{\mathbf{X}}(t,s)\beta(s)\,ds \tag{9}$$

In practice, this results in a large and often 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. 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

$$Y_i = \sum_{j=1}^{n} \beta(t_j)X_i(t_j) + \epsilon_i, \tag{10}$$

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˙introduction˙20**
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)
- 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˙introduction˙2017**
- **hsing˙theoretical˙2015**
- **ramsay˙functional˙2005**
- **horvath˙inference˙2012**
- **cai˙prediction˙2006**
- **levitin˙introduction˙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˙explorations˙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˙life˙2002**

5 Outlook

5.1 Literature

- **James.2009** (shape-restrictions)

6 Appendix

7 Bibliography