Econometrics and Statistics 1 (2017) 167–183

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

High-dimensional adaptive function-on-scalar regression

Zhaohu Fan[a], Matthew Reimherr[b,*]

[a] Departments of Industrial Engineering and Statistics, Penn State University, University Park, PA 16802, United States
[b] Department of Statistics, Penn State University, University Park, PA 16802, United States

ARTICLE INFO

Article history:
Received 5 February 2016
Revised 8 August 2016
Accepted 9 August 2016
Available online 9 November 2016

Keywords:
Variable selection
Functional regression
Oracle property

ABSTRACT

Applications of functional data with large numbers of predictors have grown precipitously in recent years, driven, in part, by rapid advances in genotyping technologies. Given the large numbers of genetic mutations encountered in genetic association studies, statistical methods which more fully exploit the underlying structure of the data are imperative for maximizing statistical power. However, there is currently very limited work in functional data with large numbers of predictors. Tools are presented for simultaneous variable selection and parameter estimation in a functional linear model with a functional outcome and a large number of scalar predictors; the technique is called AFSL for *Adaptive Function-on-Scalar Lasso*. It is demonstrated how techniques from convex analysis over Hilbert spaces can be used to establish a functional version of the oracle property for AFSL over any real separable Hilbert space, even when the number of predictors, I, is exponentially large compared to the sample size, N. AFSL is illustrated via a simulation study and data from the Childhood Asthma Management Program, CAMP, selecting those genetic mutations which are important for lung growth.

1. Introduction

Many scientific areas are faced with the challenge of extracting information from increasingly large, complex, and highly structured data sets. A great deal of modern statistical work focuses on developing tools for handling such data. Networks, high dimensional data, images, functions, surfaces, or shapes, all present data structures which are not well handled under a traditional univariate or multivariate statistical paradigm. In this paper we present a new methodology, which we call *Adaptive Function-on-Scalar Lasso*, AFSL, for analyzing highly complex functional outcomes alongside large numbers of scalar predictors. Such data is becoming increasingly common due to the prevalence of inexpensive genotyping technologies. Genome-wide association studies, GWAS, examine hundreds of thousands or millions of genetic markers, attempting to find those mutations which are associated with some outcome or phenotype. Many phenotypes of interest are now complex outcomes, such as longitudinal measurements or biomedical images.

The functional linear model, FLM, is one of the primary modeling tools in FDA. There, one assumes that the outcome is linearly related to some set of predictors. FLM are often categorized by whether the outcome, the predictor, or both is functional (Reiss et al., 2010). While the literature on FLM is now vast, Morris (2015) outlines most of the key work on low dimensional FLM. However, to date, relatively little has been done when one has a large number of predictors

* Corresponding author.
E-mail addresses: mlr36@psu.edu, mreimherr@psu.edu (M. Reimherr).

http://dx.doi.org/10.1016/j.ecosta.2016.08.001
2452-3062/

relative to the sample size, and of the work that does exist, nearly all of it is for scalar-on-function FLM, the opposite setting we consider (Fan et al., 2015; Gertheiss et al., 2013; Lian, 2013; Matsui and Konishi, 2011). For the function-on-scalar setting, we are aware of only two other works. Chen et al. (2016) consider a basis expansion approach with an MCP style penalty and fixed number of covariates. They also use a *pre-whitening* technique to exploit the within function dependence of the outcomes. Asymptotic theory is developed for a fixed number of predictors and basis functions. Barber et al. (2016) developed the Function-on-Scalar LASSO, FSL, which allows for an exponential number of predictors relative to the sample size and establishes optimal convergence rates of the estimates. In particular, it was shown that FSL achieves the same rates of convergence as in the scalar case. However, as in the scalar case, this approach leads to estimates with a non-negligible asymptotic bias due to the nature of the penalty.

To alleviate the bias problem inherent in FSL, we propose here an adaptive version, AFSL. In addition to providing a novel statistical methodology, we develop a new theoretical framework needed to establish its asymptotic properties. In particular, functional subgradients and tools from convex analysis over Hilbert spaces are needed. In contrast, theory for FSL is built entirely on *basic* inequalities and *concentration* inequalities, no theory for subgradients was required. The contributions of this paper are thus as follows (1) we define a new variable selection and estimation tool, AFSL, which alleviates the bias problems of FSL (2) we demonstrate how several tools and techniques from convex analysis can be used for functional data problems and (3) we define a functional version of the oracle property and show that AFSL achieves it. Additionally, we also go a step beyond the traditional oracle property which states that the estimates recover the correct support and are asymptotically normal, by showing that the oracle estimate and the AFSL estimate are actually asymptotically equivalent.

The remainder of the paper is organized as follows. In Section 2, we provide the necessary background material. In Section 3 we outline the AFSL framework and in Section 4 we establish the oracle property. A simulation study and an application to the Childhood Asthma Management Program, CAMP, is given in Section 5. Concluding remarks are given in Section 6. All theory is provided in the Appendix.

2. Background

Let $\mathcal{H}$ denote a real separable Hilbert space, $\langle \cdot, \cdot \rangle$ its inner product, and $\|\cdot\|$ the inner product norm. In a function-on-scalar linear model we have that

$$Y_n = \sum_{i=1}^{I} X_{ni}\beta_i^\star + \varepsilon_n = X_n^\top \beta^\star + \varepsilon_n \quad 1 \le n \le N, \tag{1}$$

where $Y_n \in \mathcal{H}$ is a functional outcome, $\beta_i^\star \in \mathcal{H}$ is a functional regression parameter, $\varepsilon_n \in \mathcal{H}$ is a functional error, and $X_{ni} \in \mathbb{R}$ is a scalar predictor. Throughout, we will use a $\star$ to denote the true data generating parameter so as to distinguish from β, which will usually represent a dummy variable or the argument of a function. The most commonly encountered space for $\mathcal{H}$ is $L^2[0, 1]$, i.e., the outcome is a function of time, though other spaces are used as well including spatial domains or product spaces (for functional panels or multivariate functional data). If one wants to incorporate smoothness assumptions of the data then one can work with Sobolev spaces or Reproducing Kernel Hilbert Spaces, RKHS. We provide a few examples here to help emphasize the functional nature of the data and highlight the wide impact of our theory and methods.

Example 1. Let $\mathcal{H} = L^2(\mathcal{D})$, where $\mathcal{D}$ is a compact subset of $\mathbb{R}^d$. Then we write the model

$$Y_n(\mathbf{s}) = \sum_{i=1}^{I} X_{ni}\beta_i^\star(\mathbf{s}) + \varepsilon_n(\mathbf{s}) \qquad \mathbf{s} \in \mathcal{D},$$

and the norm is written as

$$\|x\|^2 = \int_{\mathcal{D}} x(\mathbf{s})^2 \, d\mathbf{s}.$$

Example 2. Let $\mathcal{H} = H^{1,2}[0, 1]$, i.e., the Sobolev space of real valued functions over the unit interval with one square integrable derivative. Then we have

$$Y_n(t) = \sum_{i=1}^{I} X_{ni}\beta_i^\star(t) + \varepsilon_n(t) \qquad t \in [0, 1],$$

with the norm given by

$$\|x\|^2 = \int_0^1 x(t)^2 \, dt + \int_0^1 x'(t)^2 \, dt.$$

Example 3. Let $\mathcal{H}$, be an RKHS of functions over the unit interval with kernel operator K. Then we have

$$Y_n(t) = \sum_{i=1}^{I} X_{ni}\beta_i^\star(t) + \varepsilon_n(t),$$