Journal of Multivariate Analysis 156 (2017) 39–56

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Regularized partially functional quantile regression

Fang Yao [a,b,*], Shivon Sue-Chee [b], Fan Wang [b]

[a] Institute of Statistics, Nankai University, 94 Weijin Road, Tianjin, 300071, China

[b] Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada M5S 3G3

ARTICLE INFO

Article history:
Received 13 August 2016
Available online 10 February 2017

AMS subject classifications:
62G08
62H25
62J07

Keywords:
Functional data
Penalization
Principal components
Quantile regression

ABSTRACT

We propose a regularized partially functional quantile regression model where the response variable is scalar while the explanatory variables involve both infinite-dimensional predictor processes viewed as functional data, and high-dimensional scalar covariates. Despite extensive work focusing on functional linear models, little effort has been devoted to the development of robust methodologies that tackle the scenarios of non-normal errors. This motivates our proposal of functional quantile regression that seeks an alternative and robust solution to least squares type procedures within the partially functional regression framework. We focus on estimating and selecting the important variables in the high-dimensional covariates, which is complicated by the infinite-dimensional functional predictor. We establish the asymptotic properties of the resulting shrinkage estimator, and empirical illustrations are given by simulation and an application to a brain imaging dataset.

1. Introduction

With the advancement in technology, functional data are becoming more prevalent. The data are often observed on a dense or sparse grid of time or space. Because the number of points usually exceeds the sample size, it is natural to assume that repeated measurements correspond to smooth trajectories rather than vectors. Hence, due to their intrinsic infinite-dimensional nature, functional predictors suffer from the "curse of dimensionality" and traditional regression approaches do not suffice to solve the ill-posed inverse problem.

The functional linear model was introduced by [17] to study the effect of a functional predictor that is realized by an integral form of the regression parameter function. To date, the monograph [18] offers a comprehensive introduction to the topic of functional data analysis. To deal with the need for regularization, the trajectories and the regression parameter function are often projected unto pre-fixed basis systems such as eigenbases, wavelet or spline bases. A popular dimension reduction tool is the use of data-adaptive eigenbasis functions resulting from functional principal component (FPC) analysis, which has been studied in [19,24,26], among others. Recent theoretical properties were developed in [8,6] and applied to various functional linear model settings including longitudinal data, generalized models, additive models, polynomial models, and varying coefficient models.

It is often the case, notably in designed experiments, that one has access to several potential covariates besides functional covariates and this scenario motivates our work; this is referred to as a "mixed" or "hybrid" data setting in [18]. In linear regression for a scalar response, with both vector-valued and function-valued predictors, Kong et al. [13] consider this as a partially functional or more generally, as a semiparametric functional framework. The scalar covariates form the traditional

* Corresponding author.
E-mail address: fyao@utstat.toronto.edu (F. Yao).

http://dx.doi.org/10.1016/j.jmva.2017.02.001
0047-259X/

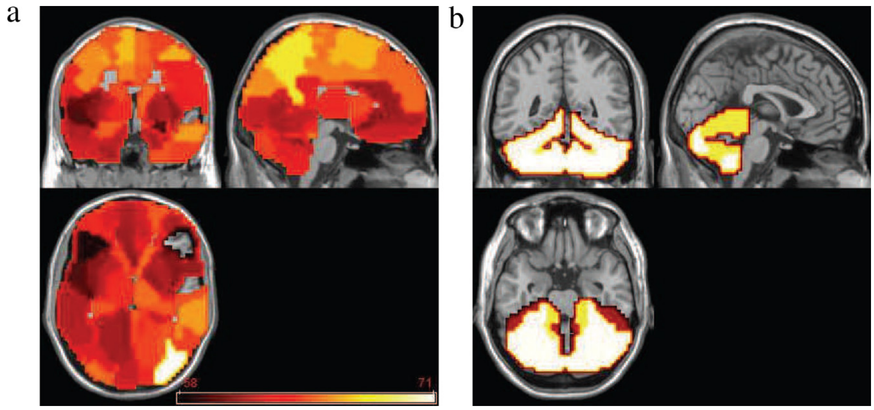

Fig. 1. (a) fMRI images of brain of an individual at the starting point of experiment. Images were taken using 3D slicer and magnified to enhance visualization. (b) Back, side and top view of the brain of a subject highlighting the region of the cerebellum.

parametric component, while the nonparametric component is associated with functional data analysis. This framework focuses on inferring the effect of important non-functional predictors while simultaneously accounting for additional information from functional predictors. It has the advantage of leading to more interpretable results than a purely functional model.

In this article, we extend partially functional linear regression into the quantile regression setting toward establishing a more flexible and robust approach. The seminal work [12] established quantile regression as an alternative to least squares regression with many desirable properties. For instance, by estimating several conditional quantile functions, compared with least squares regression, one may better account for the location, scale and shape of the response distribution. A comprehensive review was given by [11] on the topic of quantile regression; applications are often found in economics, finance, and health.

In this paper, we consider quantile regression within the partially functional framework and further consider that the number p_n of scalar predictors increases with sample size n. Various studies have considered predictors of various dimensions within linear quantile regression, such as fixed-dimensional p in [15,23,28], while [2,22] dealt with high-dimensional p and [9] addressed the case of an infinite-dimensional functional predictor variable.

Our proposal is motivated in part by the need to model health scores and economic indices which tend to be non-normal with severe skewness and outliers as in, e.g., the attention deficit hyperactivity disorder (ADHD) index we consider as the response variable in our data application. This index measures the severity of the brain disorder among the individuals. Our explanatory variables consist of both functional data, represented by functional magnetic resonance imaging (fMRI) data demonstrated in Fig. 1, and several non-functional covariates such as sex, age, diagnosis status, medication status and IQ scores. We note that the response variable, ADHD index, is heavily right-skewed even under a logarithm transformation. Hence, we propose our partially functional quantile model to accommodate for more general, non-normal distributions of the response and compare the results with that from the partially functional linear analog. Further, we present a more comprehensive description of the changes in the relationship between the response and the functional and non-functional covariates as we vary the quantile index.

To our knowledge, there has been no work covering robust estimation for partially functional model with high-dimensional scalar covariates, while [25] studied this type of quantile model with a finite number of scalar covariates that does not require penalization. This may be associated with several technical challenges. Functional data are usually not directly observed but rather we often see longitudinally repeated measurements. Hence, the need to properly recover infinite-dimensional functional data from seemingly discrete observations. Therefore, the effect of stochastic errors due to FPC estimation procedures, the unified framework of functional and non-functional predictors, the increasing dimensionality of non-functional parameters, and the non-differentiable quantile loss function together offer significant challenges.

The rest of this paper is organized as follows. We describe the proposed regularized partially functional quantile regression in Section 2, including model formulation, selection, estimation and implementation procedures. Theoretical properties are presented in Section 3, and empirical performance is shown in Section 4 via simulations. An application to the ADHD-200 fMRI data is given in Section 5. Regularity conditions, auxiliary lemmas and proofs of the main theorems are deferred to the Appendix.

2. Proposed methodology

Suppose that we observe a random sample $(X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n)$ from a random triple (X, Z, Y), where the predictor X is a twice continuously differentiable and squared integrable random process supported on a closed interval $\mathcal{T}$, Z is a p-dimensional vector covariate, and the response Y is scalar. This sets up a partially functional regression framework