Neurocomputing 195 (2016) 74–87

Contents lists available at ScienceDirect

Neurocomputing

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

Partial functional linear quantile regression for neuroimaging data analysis ☆

Dengdeng Yu, Linglong Kong *, Ivan Mizera

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2G1

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
13 June 2015
Accepted 14 August 2015
Available online 4 February 2016

Keywords:
Functional linear quantile regression
Partial quantile covariance
PQR basis
SIMPQR
ADHD
ADNI

ABSTRACT

We propose a prediction procedure for the functional linear quantile regression model by using partial quantile covariance techniques and develop a simple partial quantile regression (SIMPQR) algorithm to efficiently extract partial quantile regression (PQR) basis for estimating functional coefficients. We further extend our partial quantile covariance techniques to functional composite quantile regression (CQR) defining partial composite quantile covariance. There are three major contributions. (1) We define partial quantile covariance between two scalar variables through linear quantile regression. We compute PQR basis by sequentially maximizing the partial quantile covariance between the response and projections of functional covariates. (2) In order to efficiently extract PQR basis, we develop a SIMPQR algorithm analog to simple partial least squares (SIMPLS). (3) Under the homoscedasticity assumption, we extend our techniques to partial composite quantile covariance and use it to find the partial composite quantile regression (PCQR) basis. The SIMPQR algorithm is then modified to obtain the SIMPCQR algorithm. Two simulation studies show the superiority of our proposed methods. Two real data from ADHD-200 sample and ADNI are analyzed using our proposed methods.

1. Introduction

Nowadays, there is great need in the analysis of complex neuroimaging data obtained from various cross-sectional and clustered neuroimaging studies. These neuroimaging studies are essential to advancing our understanding of the neural development of neuropsychiatric and neurodegenerative disorders, substance use disorders, the normal brain and the interactive effects of environmental and genetic factors on brain structure and function. Such large imaging studies include the ADNI (Alzheimer's Disease Neuroimaging Initiative), the longitudinal magnetic resonance imaging (MRI) study of schizophrenia, autism, and attention deficit hyperactivity disorder (ADHD), the NIH human connectome project, among many others. Neuroimaging studies usually collect structural, neurochemical, and functional images over both time and space [15,16,33]. These structural, neurochemical, and functional imaging modalities include computed axial tomography (CT), diffusion tensor imaging (DTI), functional magnetic resonance imaging (fMRI), magnetic resonance imaging (MRI), magnetic resonance spectroscopy (MRS), positron emission tomography (PET), single photon emission tomography (SPECT), electroencephalography (EEG), and magnetoencephalography (MEG), among many others. For instance, by using anatomical MRI, various measures of the morphology of the cortical and subcortical structures (e.g., hippocampus) are extracted to understand neuroanatomical differences in brain structure across different populations [14,36]. In DTI, various diffusion properties and fiber tracts are extracted for quantitative assessment of anatomical connectivity across different populations [4,48–50]. Functional images, such as resting-state functional MRI (rsfMRI), have been widely used in behavioral and cognitive neuroscience to understand functional segregation and integration of different brain regions across different populations [21,34].

A common feature of many imaging techniques is that massive functional data are observed/calculated at the same design points, such as time for functional images (e.g., PET and fMRI) and arclength for structure imaging (e.g. DTI). As an illustration, we present two smoothed functional data that we encounter in neuroimaging studies. First, we consider the BOLD rsfMRI signal, which is based on hemodynamic responses secondary to resting-state. We plot the estimated hemodynamic response functions (HRF) with 172 time courses from 20 randomly selected children

☆*The Alzheimer's Disease Neuroimaging Initiative*: Part of the data used in preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

* Corresponding author.
E-mail address: lkong@ualberta.ca (L. Kong).

http://dx.doi.org/10.1016/j.neucom.2015.08.116
0925-2312/

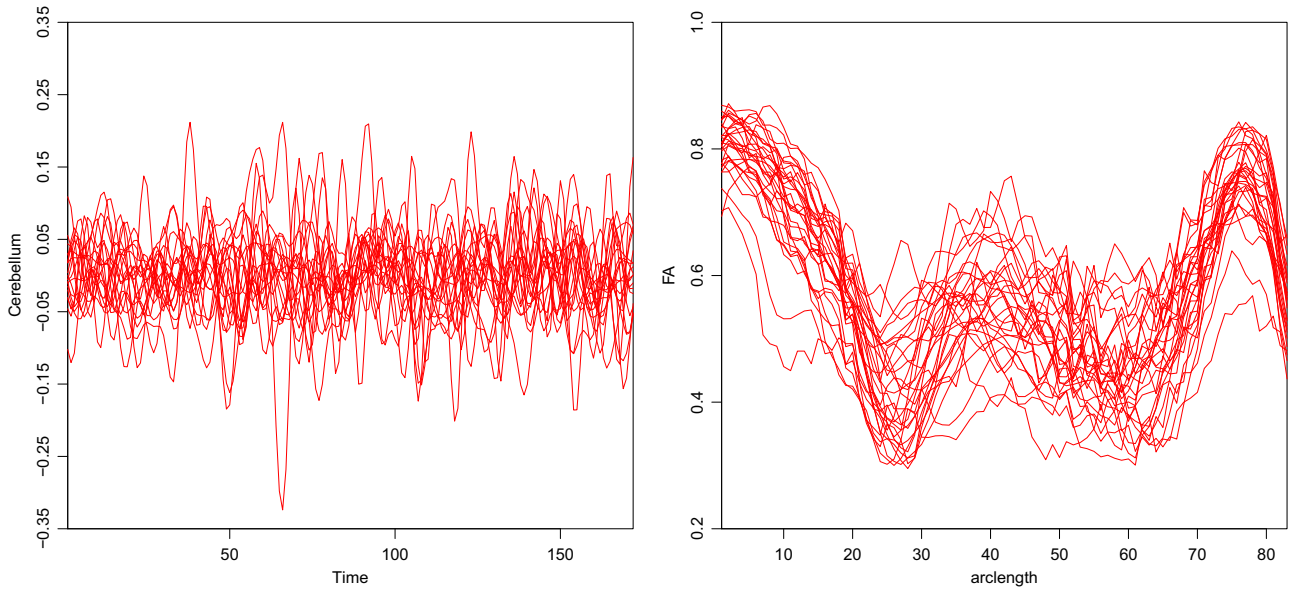

Fig. 1. Representative functional neuroimaging data: (left) the estimated hemodynamic response functions (HRF) corresponding to resting-state from 20 children at NYU from the ADHD-200 Sample Initiative Project and (right) fractional anisotropy (FA) along the midsagittal corpus callosum (CC) skeleton from 30 randomly selected subjects from the NIH Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

at a selected region of interest (ROI) of Anatomical Automatic Labeling (AAL) atlas [42] from the New York University (NYU) Child Study Center from the ADHD-200 Sample Initiative Project. Although the canonical form of the HRF is often used, when applying rsfMRI in a clinical population with possibly altered hemodynamic responses (Fig. 1 (a)), using the subject's own HRF in rsfMRI data analysis may be advantageous because HRF variability is greater across subjects than across brain regions within a subject [1,31]. We are particularly interested in delineating the structure of the variability of the HRF and their capacity of predicting ADHD index with a set of covariates of interest, such as diagnostic group [30]. Secondly, we plot one diffusion property, called fractional anisotropy (FA), measured at 83 grid points along the midsagittal corpus callosum (CC) skeleton (Fig. 1 (b)) from 30 randomly selected infants from the NIH Alzheimer's Disease Neuroimaging Initiative (ADNI) study. The corpus callosum (CC) is the largest fiber tract in the human brain and is a topographically organized structure. It is responsible for much of the communication between the two hemispheres and connects homologous areas in the two cerebral hemispheres. Scientists are particularly interested in delineating the structure of the variability of these functional FA data and their prediction ability on mini-mental state examination (MMSE) with a set of covariates of interest, such as genetic information. MMSE is one of the most widely used screening tests on Alzheimer's Disease to provide brief and objective measures of cognitive functioning [41]. We will systematically investigate these two prediction problems using functional imaging data over time or space in Section 7 after we develop our methodology.

A functional linear regression model, where the responses such as the neurological or clinical outcomes (e.g. ADHD index or MMSE) are modeled by a set of scalar covariates and functional covariates of interest (e.g. HRF along time courses or FA along arclength), is a powerful statistical tool for addressing these scientific questions [17,18,48,49]. In particular, denoting the neurological or clinical outcome of the i-th subject by y_i, $i=1,\ldots,n$, the functional linear regression model is of the form

$$y_i=\alpha+\boldsymbol{x}_i^T\boldsymbol{\beta}+\int_0^1 \boldsymbol{z}_i^T(t)\boldsymbol{\gamma}(t)\,dt+\epsilon_i, \tag{1}$$

where α is the intercept, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$ is a $p\times 1$ vector of coefficients, $\boldsymbol{x}_i=(x_{i1},\ldots,x_{ip})^T$ is a $p\times 1$ vector of scalar covariates of interest, $\boldsymbol{\gamma}(t)=(\gamma_1(t),\ldots,\gamma_q(t))^T$ is a $q\times 1$ vector of coefficient functions of t, $\boldsymbol{z}_i(t)=(z_{i1}(t),\ldots,z_{iq}(t))^T$ is a $q\times 1$ vector of functional covariates, and ϵ_i is a random error. It is usually assumed that ϵ_i is independent and identical copy of normal distribution with zero mean and variance σ^2. For simplicity, we let $t\in[0,1]$. Model (1) is a generalization of the classical linear regression model corresponding to the case $\boldsymbol{\gamma}(t)$ is a constant. If it is not constant, the contributions of $\boldsymbol{z}_i(t)$ characterized by $\boldsymbol{\gamma}(t)$ change in terms of t. The model has been well studied and applied in many fields including neuroimaging data analysis [2,23,29,37]. To facilitate the estimation of $\boldsymbol{\gamma}(t)$, we usually require that it satisfies certain smoothness conditions and restrict it onto a functional space. For example, we may require that its second derivative exists and that the square of $\gamma(t)$ is integrable, that is, $\boldsymbol{\gamma}(t)\in L_2[0,1]$. Even in such a case, the estimation is still an infinite-dimensional problem.

The common practice is to project $\boldsymbol{\gamma}(t)$ into a functional space with a finite functional basis. There are three major methods to choose the functional basis: general basis, functional principal component (fPC) basis, and partial least square (PLS) basis. There are various options on the selection of general basis, for example B-spline basis [6,9], wavelet basis [45] and so on. In order to provide a good approximation of the functional coefficients, a large number of basis should be chosen. However, this may cause overfitting of the model and to remedy that various penalty methods have been proposed [10,44]. The fPC method has been extensively studied [19,27] where the fPC of $\boldsymbol{z}_i(t)$ serve as the basis. Its generalization to the reproducing kernel Hilbert space (RKHS) was proposed by Cai and Yuan [7,43] who also studied its minimax rates. Although fPC basis is more data-adapted than the general basis as they use the information of functional covariates and the formed space can explain most of the variation of $\boldsymbol{z}_i(t)$, it is not necessary that all the fPC basis will contribute to the variation of the responses. Therefore, another appealing choice is the PLS basis which uses both the information of functional covariates and the responses. The PLS basis uses the linear projects of $\boldsymbol{z}_i(t)$ which best predicts the responses [12].

An alternative to model (1) is the functional linear quantile regression where the conditional quantiles of the responses are modeled by a set of scalar covariates and functional covariates. There are at least three advantages to use conditional quantiles