Journal of Multivariate Analysis 165 (2018) 279–295

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Simultaneous inference for the mean of repeated functional data

Guanqun Cao [a], Li Wang [b,*]

[a] Department of Mathematics and Statistics, Auburn University, United States
[b] Department of Statistics and the Statistical Laboratory, Iowa State University, United States

ARTICLE INFO

Article history:
Received 29 May 2017
Available online 9 February 2018

AMS 2000 subject classifications:
primary 62M10
secondary 62G08

Keywords:
B-spline
Confidence band
Functional data
Repeated measures
Semiparametric efficiency

ABSTRACT

Motivated by recent works studying the longitudinal diffusion tensor imaging (DTI) studies, we develop a novel procedure to construct simultaneous confidence bands for mean functions of repeatedly observed functional data. A fully nonparametric method is proposed to estimate the mean function and variance–covariance function of the repeated trajectories via polynomial spline smoothing. The proposed confidence bands are shown to be asymptotically correct by taking into account the correlation of trajectories within subjects. The procedure is also extended to the two-sample case in which we focus on comparing the mean functions from two populations of functional data. We show the finite-sample properties of the proposed confidence bands by simulation studies, and compare the performance of our approach with the "naive" method that assumes the independence within the repeatedly observed trajectories. The proposed method is applied to the DTI study.

1. Introduction

With modern technological progress in measuring devices, sophisticated data are now easy to collect. These data are often sets of functions such as curves, images or shapes, whose high-dimensional and correlated features impose tremendous challenges on conventional statistical studies. Emerging as a promising field, functional data analysis (FDA), which deals with the analysis of curves, has recently undergone intense development. The interested reader is referred to Ramsay and Silverman [16] for a general introduction of FDA.

In this work, we focus on situations where curves are repeatedly recorded for each subject, e.g., mortality data [6] in which age-specific lifetables are collected over years for various countries, and electroencephalography (EEG) data [9] observed for patients at each visit. Such dependent types of curves or images now commonly arise in diverse fields including climatology, demography, economics, epidemiology, and finance.

Our work is motivated by a longitudinal neuroimaging study containing repeated functional measurements derived from diffusion tensor imaging (DTI); for a description, see [11,12]. DTI is a magnetic resonance imaging technique which provides different measures of water diffusivity along brain white matter tracts; its use is instrumental, especially in diseases that affect the brain white matter tissue such as multiple-sclerosis (MS); see, e.g., [1]. In this study, DTI brain scans are recorded for many multiple-sclerosis (MS) patients to assess the effect of neurodegeneration on disability. At each visit, fractional anisotropy (FA) was determined via DTI along the corpus callosum (CCA). One objective here is to better understand the

* Corresponding author.
E-mail address: lilywang@iastate.edu (L. Wang).

https://doi.org/10.1016/j.jmva.2018.02.001

demyelination process via its FA proxy and investigate possible differences therein between female and male patients. Point-wise confidence intervals via estimation ± 2 point-wise standard errors are provided in [17]. It is unclear, however, what is the performance of global inference on the true underlying mean profile.

In this paper, we develop simultaneous inference for the mean of repeated functional data. Our approach can handle the within-subject correlation and provide global inference, which are the key advantages of our approach over available FDA methods. There have been some recent attempts to study such repeated functional data in various settings. For example, the importance of models for dependent functional data has been recognized in [7,9], in which the emphasis has been on a general hierarchical model. Chen and Müller [6] proposed a flexible longitudinally observed functional model and provided consistency results and asymptotic convergence rates for the estimated model components. Zhu et al. [23] established the uniform convergence rate and confidence band for each estimated individual effect curve in multivariate varying coefficient models.

Simultaneous confidence bands (SCBs) are an important tool to address the variability in the unknown function and to develop global test statistics for general hypothesis testing problems. In Wang et al. [19,20], smooth SCBs are developed for the cumulative distribution functions. Gu and Yang [14] constructed SCBs for the link function in a single-index model based on the oracally efficient kernel estimator. It is of particular interest in FDA to construct SCBs for mean functions. For example, Bunea et al. [2] proposed an asymptotically conservative confidence set for the mean function of Gaussian functional data. Song et al. [18] proposed an asymptotically correct SCBs for dense functional data using local linear smoothing. Recently, polynomial splines have found successful applications in SCB construction. Ma et al. [15] suggested spline SCBs for mean functions of sparse functional data based on polynomial spline smoothing. Gu et al. [13] investigated a varying coefficient regression model for sparse functional data and proposed simultaneous confidence corridors for the coefficient functions. Cao et al. [4,5] provided SCBs for mean and derivative functions of dense functional data, respectively.

In this paper, we derive SCBs for mean functions when curves are repeatedly recorded for each subject. Existing methodologies for constructing SCBs in FDA often assume the independence of trajectories within each subject. Thus, the within-subject effect is not reflected by the traditional covariance functions of the mean curve. We are unaware of any methodology that provides exact SCBs for mean curves of repeatedly observed functional data. In this work, we use polynomial splines to approximate the mean and covariance functions in the construction of the SCBs. We show that the proposed spline SCBs are asymptotically correct and semiparametrically efficient in the sense that they are asymptotically the same as if all random trajectories were observed entirely and without errors as in [5]. We further consider two-sample inference for dependent functional data and extend our SCB construction procedure to a two-sample problem to test whether the mean functions from two groups are different.

The dependence within the repeatedly observed curves adds extra difficulty for model implementation, e.g., the estimation of within-subjects correlation. Misspecification of the correlation structure may lead to some efficiency loss. To tackle this issue, it is desirable to make the structure as model-free as it can be, and nonparametric modeling is particularly useful in this sense. In this paper we propose to estimate the variance–covariance functions nonparametrically. Our Monte Carlo results show that the proposed bands have much more accurate coverage rates of the true function compared to the "naive" method that ignores the within-subject dependence.

The paper is organized as follows. Section 2 states the model and introduces the estimates of mean functions for repeated functional data. Section 3.1 describes the asymptotic distribution of the estimators in the framework of allowing unknown dependence of the trajectories within subjects. Using this asymptotic result, we construct SCBs for mean functions. Section 3.2 develops the SCBs to study the difference of mean functions from two populations. Section 4 discusses how to estimate the components in the proposed bands. A simulation study is presented in Section 5. Section 6 contains applications of our method to a diffusion tensor imaging data. Section 7 gives the concluding remarks. Further insights into the error structure of spline estimators and technical proofs are collected in the Appendix.

2. The model and estimates

2.1. Modeling repeated functional measurements

We consider data $\{X_{ij}(s) : s \in \mathcal{X}\}$, $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m_i\}$, where X_{ij} is a repeated random curve on the compact interval $\mathcal{X}$, i is the subject index, and j is the repeated trajectory index for the ith subject. Assume that for all $j \in \{1, \ldots, m_i\}$, X_{ij} are iid copies of the L_2 process X_j defined on $[0, 1]$, with mean function defined, for all $s \in [0, 1]$, by $\mu(s) = \mathrm{E}\{X_{ij}(s)\}$.

For the ith subject one has the Karhunen–Loève representation of the process of $X_{ij}(s)$, i.e., $X_{ij}(s) = \mu(s) + \sum_{k=1}^{\infty} \xi_{ijk}\phi_{jk}(s)$, where the random coefficients ξ_{ijk}s are referred to as the (jk)th functional principal component (FPC) scores of the ith subject. For each fixed (i, j), the ξ_{ijk}s are uncorrelated with mean 0 and variance 1. For notational convenience, let $\phi_{jk} = \sqrt{\lambda_{jk}}\, \psi_{jk}$; then λ_{jk} and ψ_{jk} are the eigenvalues and eigenfunctions of the covariance operator with kernel $G_{jj}(s, t) = \mathrm{cov}\{X_{1j}(s), X_{1j}(t)\}$, respectively. Although the sequences $\{\lambda_{jk}\}_{j,k=1}^{m_i,\infty}$, $\{\phi_{jk}\}_{j,k=1}^{m_i,\infty}$ and the random coefficients ξ_{ijk}s exist; however, they are unknown or unobservable.

Let $\mathbf{Y}_i(s) = (Y_{i1}(s), Y_{i2}(s), \ldots, Y_{im_i}(s))^\top$ for all $i \in \{1, \ldots, n\}$, and assume $Y_{ij}(s) = X_{ij}(s) + \varepsilon_{ij}(s)$, where $\varepsilon_{ij}(s)$ are mean zero measurement errors. Suppose $X_{ij}(s) = \mu(s) + \eta_{ij}(s)$, where $\eta_{ij}(s)$ characterizes individual curve variations from $\mu(s)$. Denote $\boldsymbol{\varepsilon}_i(s) = (\varepsilon_{i1}(s), \ldots, \varepsilon_{im_i}(s))^\top$ and $\boldsymbol{\eta}_i(s) = (\eta_{i1}(s), \ldots, \eta_{im_i}(s))^\top$. Suppose $\boldsymbol{\varepsilon}_i(s)$ and $\boldsymbol{\eta}_i(s)$ are mutually independent. Moreover, assume that $\boldsymbol{\eta}_i(s)$ and $\boldsymbol{\varepsilon}_i(s)$ are iid copies of stochastic processes with mean vector $\mathbf{0}$ and covariance functions