Computational Statistics and Data Analysis 106 (2017) 153–164

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimating functional linear mixed-effects regression models

Baisen Liu [a], Liangliang Wang [b], Jiguo Cao [b,*]

[a] School of Statistics, Dongbei University of Finance and Economics, Dalian 116025, China

[b] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A1S6, Canada

ARTICLE INFO

Article history:
Received 5 January 2016
Received in revised form 26 May 2016
Accepted 19 September 2016
Available online 7 October 2016

Keywords:
EM algorithm
Functional linear regression
Smoothing spline
Random effects model

ABSTRACT

A new functional linear mixed model is proposed to investigate the impact of functional predictors on a scalar response when repeated measurements are available on multiple subjects. The advantage of the proposed model is that under the proposed model, each subject has both individual scalar covariate effects and individual functional effects over time, while it shares the common population scalar covariate effects and the common population slope functions. A smoothing spline method is proposed to estimate the population fixed and random slope functions, and a REML-based EM algorithm is developed to estimate fixed effects and variance parameters for random effects. Simulation studies illustrate that for finite samples the proposed estimation method can provide accurate estimates for the functional linear mixed-effects model. The proposed model is applied to investigate the effect of daily ozone concentration on annual nonaccidental mortality rates and also to study the effect of daily temperature on annual precipitation.

1. Introduction

When a random variable is observed at multiple time points or spatial locations, the data can be viewed as a function of time or spatial location. This type of data is generally called functional data (Ramsay and Silverman, 2005). Now in the big data era, functional data analysis (FDA) has become very attractive in statistical methodology and applied data analysis. Functional linear models (FLMs), introduced by Ramsay and Dalzell (1991), are some of the most popular models in FDA. FLMs can be used to model the relationship between functional variables and to predict a scalar response from functional covariates. Developments in modern technology have allowed FLMs to be applied to model functional data in many fields such as economics, medicine, environment, and climate [see for instance, Ramsay and Silverman, 2002, 2005, and Ferraty and Vieu, 2006, for several case studies].

The properties of FLMs have been thoroughly examined in the literature. For example, Yao et al. (2005) studied FLMs for sparse longitudinal data and suggested a nonparametric estimation method based on functional principal components analysis (FPCA). Their proposed functional regression approach is flexible enough to allow sparse measurements of functional predictors and response. Cai and Hall (2006) discussed the prediction problem in FLMs based on the FPCA technique. Crambes et al. (2009) proposed a smoothing spline estimator for the functional slope parameter, and extended this estimator to covariates with measurement-errors. Yuan and Cai (2010) suggested a smoothness regularization method for estimating FLMs based on the reproducing kernel Hilbert space (RKHS) approach. They provided a unified treatment for both the prediction and estimation problems by developing a tool that relys on simultaneous diagonalization of two positive-definite kernels. Wu et al. (2010) proposed a varying-coefficient FLM which allows for the slope to be modeled

* Correspondence to: Room SC K10545, 8888 University Drive, Burnaby, BC, Canada V5A1S6.
E-mail address: jiguo_cao@sfu.ca (J. Cao).

http://dx.doi.org/10.1016/j.csda.2016.09.009

as a dependent function of additional scalar covariates. Gertheiss et al. (2013) extended the classical functional principal components regression (FPCR) by developing a longitudinal FPCR that allows for different effects of subject-specific trends in curves and for visit-specific deviations from that trend in longitudinal functional data. Scheipl et al. (2015) proposed a flexible functional additive mixed model that incorporates linear and nonlinear effects of functional and scalar covariates and allowed for flexible correlation structures of data. A systematic review on FLMs can be found in Morris (2015).

One conventional FLM involves linking a scalar response variable $Y_j, j = 1, \ldots, m$, to a functional predictor $X_j(t)$ through the following model:

$$Y_j = \alpha + \int_S \beta(t) X_j(t) \mathrm{d}t + \epsilon_j, \tag{1}$$

where α is the intercept, $\beta(t)$ is a smooth slope function, and the ϵ_j's are independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ_ϵ^2. In this model, S is often assumed to be a compact subset of an Euclidean space such as $[0, 1]$. The slope function $\beta(t)$ represents the accumulative effect of the functional covariate $X_j(t)$ on the scalar response Y_j.

For purposes of illustration, we take the air pollution data as an example. This data is from the R package NMMAPSdata (Peng and Welty, 2004), which catalogues air pollution, weather, and mortality data for American cities from 1987 to 2000. Our aim is to investigate the impact of daily ozone concentration on nonaccidental mortality rates. The scalar response Y_{ij} is the logarithm of annual nonaccidental mortality rates in the jth year at the ith city, and the functional predictor $X_{ij}(t)$ is daily ozone concentration. In a preliminary analysis, we performed the classical FLM (1) on each individual city and found that there was dramatic variation in the estimate $\hat{\beta}(t)$ among different cities. This indicates that the effect of daily ozone concentration on the annual nonaccidental mortality rates is not the same in different cities. Therefore, it may not be appropriate to pool the data for all US cities and provide a single estimate for the average effect of daily ozone on the annual nonaccidental mortality rates. On the other hand, we may not fully exploit the information available in the data if we fit separate functional linear models for each city.

To address this dilemma, we generalize the FLM (1) to allow for the incorporation of random effects in the slope function. We call this new model the functional linear mixed-effects model (FLMM). Assume that we repeatedly observe a distinct functional predictor and scalar outcome for each subject over several visits. Then the observed data has the structure $\{Y_{ij}, \mathbf{X}_{ij}(t), \mathbf{W}_{ij}, \mathbf{Z}_{ij}\}, i = 1, \ldots, n, j = 1, \ldots, m_i$, where Y_{ij} is the jth repeated measurement of the scalar response for the ith subject, $\mathbf{W}_{ij}$ and $\mathbf{Z}_{ij}$ are vectors of scalar covariates, and $\mathbf{X}_{ij}(t) = (X_{ij1}(t), X_{ij2}(t), \ldots, X_{ijd}(t))$ are the corresponding functional predictors. The functional linear mixed-effects model can be expressed as follows:

$$Y_{ij} = \mathbf{W}_{ij}'\boldsymbol{\alpha} + \mathbf{Z}_{ij}'\boldsymbol{\gamma}_i + \sum_{\ell=1}^{d} \int_S [\beta_\ell(t) + b_{i\ell}(t)] X_{ij\ell}(t) \mathrm{d}t + \epsilon_{ij}, \tag{2}$$

where $\boldsymbol{\alpha}$ is a p-dimensional vector of the fixed effect, $\boldsymbol{\gamma}_i$ is the q-dimensional vector of random effect of scalar covariates, $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_d(t))'$ represents the population effect of $\mathbf{X}_{ij}(t)$ on Y_{ij}, $\mathbf{b}_i(t) = (b_{i1}(t), \ldots, b_{id}(t))'$ stands for the random effect of $\mathbf{X}_{ij}(t)$ on Y_{ij} for the ith subject, and ϵ_{ij} is the i.i.d. random variable with mean 0 and variance σ_ϵ^2. In this article, we assume that $\boldsymbol{\gamma}_i \sim N(0, \sigma_\epsilon^2 \boldsymbol{\Psi})$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, and $b_{i\ell}(t)$ follows a Gaussian stochastic process with mean 0 and covariance function $\gamma_\ell(s, t)$, that is, $b_{i\ell}(t) \sim GP(0, \gamma_\ell(s, t))$. We also assume that $\boldsymbol{\gamma}_i$, ϵ_{ij}, $\mathbf{b}_i(t)$, and $\mathbf{X}_{ij}(t)$ are mutually independent.

The functional linear mixed-effects model above is attractive, because it can estimate the population effect and random effect of the functional predictor $\mathbf{X}(t)$ (e.g., the daily temperature and ozone concentrations) on the scalar response Y (e.g., the annual nonaccidental deaths), as well as the population effect $\boldsymbol{\alpha}$ and random effect $\boldsymbol{\gamma}_i$. The application of the proposed functional linear mixed-effects model to the air pollution problem is not unique; many similar applications can be found in environmental and biological problems.

The proposed functional linear mixed-effect model (2) is different from the following functional mixed model (Goldsmith et al., 2011, 2012):

$$Y_{ij} = \mathbf{Z}_i \mathbf{b}_i + \int_S \beta(t) X_{ij}(t) \mathrm{d}t + \epsilon_{ij}, \tag{3}$$

where $\mathbf{b}_i \sim N(\mathbf{0}, \sigma_{\mathbf{b}}^2 \mathbf{I})$ accounts for the correlation in the repeated outcomes for the ith subject. The primary distinction between models (2) and (3) is that the subject-specific random effect $\mathbf{b}_i$ in (3) remains constant across visits, while the random effect $\mathbf{b}_i(t)$ in (2) allows this effect to vary with time. By including the random effect $\mathbf{b}_i(t)$ in (2), this model can characterize trends in the effects of functional predictors on scalar outcomes for different subjects.

Many nonparametric smoothers used for FLMs can be applied to fit model (2). In this article, we employ the smoothing spline method (Ramsay and Silverman, 2005) to estimate $\beta_\ell(t)$ and $b_{i\ell}(t)$ in (2). Next, we transform model (2) into a linear mixed-effects model (LMM). Finally, we propose an REML-based EM algorithm to fit the LMM, the efficiency of which is illustrated by examples.

The remainder of this article is organized as follows. Section 2 introduces a smoothing spline method to estimate the above functional linear mixed-effects model. Section 3 implements simulations to evaluate the finite sample performance of the smoothing spline method. The functional linear mixed-effects model is demonstrated through two real applications in Section 4. Conclusions are given in Section 5.