Computational Statistics and Data Analysis 105 (2017) 46–52

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Short communication

A note on modeling sparse exponential-family functional response curves

Jan Gertheiss [a,*], Jeff Goldsmith [b], Ana-Maria Staicu [c]

[a] Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, Erzstr. 1, 38678 Clausthal-Zellerfeld, Germany

[b] Department of Biostatistics, Columbia University, NY, USA

[c] Department of Statistics, North Carolina State University, Raleigh, USA

ARTICLE INFO

Article history:
Received 1 November 2015
Received in revised form 14 July 2016
Accepted 15 July 2016
Available online 21 July 2016

Keywords:
Binomial data
Functional principal components
Longitudinal data
Mixed models
Smoothing
Sparse sampling design

ABSTRACT

Non-Gaussian functional data are considered and modeling through functional principal components analysis (FPCA) is discussed. The direct extension of popular FPCA techniques to the generalized case incorrectly uses a marginal mean estimate for a model that has an inherently conditional interpretation, and thus leads to biased estimates of population and subject-level effects. The methods proposed address this shortcoming by using either a two-stage or joint estimation strategy. The performance of all methods is compared numerically in simulations. An application to ambulatory heart rate monitoring is used to further illustrate the distinctions between approaches.

1. Introduction

Standard methods to analyze repeatedly observed exponential family data can be separated into two main categories: *marginal models*, or population-average models, and *conditional models*, also known as mixed effects or subject-specific models. The former models focus on inference for population-level effects, and are traditionally estimated using a generalized estimating equations (GEE) framework (Liang and Zeger, 1986; Zeger and Liang, 1986; Liang and Zeger, 1995; Fitzmaurice et al., 1993). In this note, we focus on the latter framework, which focuses on the within-subject associations, because this perspective is most commonly employed in functional data analysis (FDA). The conditional model framework is particularly attractive in FDA because it models the dependence inherent in the functional observation for a subject as the realization of a latent random process, and this realization is often a quantity of interest. Unlike the common approach used in a longitudinal data analysis, which assumes parametric subject-specific effects in a mixed model, the dependence of the latent process in a functional data analysis is assumed unknown.

Functional data are commonly defined as observations on subjects that one can imagine as arising from the evaluation of a subject-specific real-valued curve at a finite grid of points. In many cases, however, the grid is irregular across subjects or sparse at the subject level. There has been tremendous interest in the analysis of functional data during the past few decades; see, e.g., Ramsay and Silverman (2005) or Ferraty and Vieu (2006) for book-length overviews, and Sørensen et al. (2013) for a brief introduction. One popular technique, used to describe the variability in a sample of curves, is functional principal

* Corresponding author. Fax: +49 5323 72 2304.
E-mail addresses: jan.gertheiss@tu-clausthal.de (J. Gertheiss), jeff.goldsmith@columbia.edu (J. Goldsmith), astaicu@ncsu.edu (A.-M. Staicu).

http://dx.doi.org/10.1016/j.csda.2016.07.010
0167-9473/

component analysis (FPCA). The conceptual FPCA model for real-valued response curves $Y_i(t)$ for $t \in \mathcal{T}$ with smooth mean function $\mu(t)$ and covariance function $\Sigma(t, t')$ is

$$Y_i(t) = \mu(t) + \sum_{k\geq 1} \xi_{ik}\psi_k(t) + \epsilon_i(t), \tag{1}$$

where $\xi_{ik} \sim (0, \lambda_k)$ are subject-specific zero-mean loadings that are uncorrelated and have variance λ_k, $\{\lambda_k, \psi_k(t)\}$ is the pair of eigenvalue/eigenfunction of the covariance $\Sigma(\cdot, \cdot)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and $\epsilon_i(\cdot)$ is some zero-mean white noise process. This is commonly known as the Karhunen–Loève (KL) expansion and is a very common approach to model functional data (see, e.g., Di et al., 2009; Greven et al., 2010; Jacques and Preda, 2014). In practice, the sum in (1) is truncated such that $E[Y_i(t)|\xi_i] = \mu(t) + \sum_{k=1}^{M} \xi_{ik}\psi_k(t)$, where M is a finite truncation.

Model (1) decomposes curves $Y_i(t)$ using shared basis functions $\psi_k(\cdot)$ and subject-specific scores ξ_{ik}, and is very similar to a mixed model: indeed, setting $M = 1$ and taking $\psi_k(t) = 1, \forall t$ yields the well-known random intercept model. However, FPCA describes the directions of variation that appear in the sample and gives a parsimonious decomposition of the complex variance structure $\Sigma(\cdot, \cdot)$. For sparse data in particular, FPCA borrows strength across subjects to estimate major patterns of variation and improve estimation of subject-specific curves (Yao et al., 2005). Estimation techniques for model (1) have been proposed and studied for dense grids, irregular observations, and sparsely observed functional data (see Ramsay and Silverman, 2005; Yao et al., 2005, and many others). A popular algorithm for FPCA, for both dense and sparse data, consists of the following main steps: (i) estimate the marginal mean by assuming independence across subjects and across grid points within a subject; (ii) estimate the marginal covariance of subject-specific deviations, again by assuming independence, and take its spectral decomposition to obtain estimates of the eigenfunctions $\psi_k(\cdot)$ and eigenvalues λ_k; and (iii) conditioning on the estimated mean function and eigenfunctions, estimate the scores ξ_{ik} in a mixed model.

Analyzing repeated exponential-family outcomes using functional data approaches is currently an area of intensive research (Hall et al., 2008; Chen et al., 2013; Serban et al., 2013; Huang et al., 2014; Gertheiss et al., 2015; Goldsmith et al., 2015; Scheipl et al., 2016). Hall et al. (2008) extended model (1) to handle non-Gaussian functional data. Here, only the *latent* process is assumed to be Gaussian. Their proposed generalized FPCA (GFPCA) model is

$$E[Y_i(t)|\xi_i] = h\left\{\mu(t) + \sum_{k=1}^{M} \xi_{ik}\psi_k(t)\right\}, \tag{2}$$

with a known response function $h(\cdot)$; more recently, this model framework has been extended to account for rare events (Serban et al., 2013) and repeated functional observations on each subject (Goldsmith et al., 2015). As in model (1), it is assumed that, conditional on the subject-specific scores ξ_{ik}, the responses $Y_i(t)$'s are independent over t. The estimation method proposed by Hall et al. for model (2) directly extends the ideas from Gaussian response FPCA. However, for non-Gaussian curves, using a marginal approach to estimate the mean and covariance in what is inherently a conditional model, and then estimating subject-specific effects based on this mean and covariance, results in poor performance: the marginal mean is biased for the mean of the specified model and, as a consequence, the estimation of basis functions and subject effects is affected negatively. This is not an issue for model (1), of course, because the marginal and conditional mean are the same.

To gain more intuition, consider binary response data arising from model (2) using a logistic link function. In this case, the marginal mean $\alpha(t) = E[Y_i(t)]$ does not equal $h(\mu(t))$ if $\mu(t) \neq 0$. More specifically (see also the Appendix), for some $t_0 \in \mathcal{T}$,

$$\alpha(t_0) < h(\mu(t_0)), \quad \text{if } \mu(t_0) > 0, \quad \text{and} \quad \alpha(t_0) > h(\mu(t_0)), \quad \text{if } \mu(t_0) < 0. \tag{3}$$

The amount of bias introduced by using a marginal estimate of the conditional mean depends on the amount of variability in the latent subject-specific process.

Example. To illustrate this point, Fig. 1 shows two simulated samples of curves generated from model (2) with $M = 1$, $\mu(t) = 2\sin(2\pi t)$ and $\psi_1(t) = \sin(2\pi t)$; in one sample of curves $\lambda_1 = 1$, and in the other $\lambda_1 = 4$. The left panel shows the curves on the logit scale, on which the data are generated and naturally interpreted; the right panel shows these transformed to the probability scale, along with the marginal means of both samples. Although the samples share a mean function $\mu(t)$ on the logit scale and share a conditional mean (given $\xi = 0$) on the probability scale, both marginal means are biased estimates of $h(\mu(t))$, and a larger bias is observed for the more variable sample. Indeed, when deriving their marginal approach to estimate parameters in model (2), Hall et al. (2008) assumed the variation of the latent process around mean $\mu(t)$ to be relatively small. This assumption, however, is easily violated in practice, and Hall et al. do not provide suggestions for this case.

In this note, we argue that the GFPCA model (2) is correctly understood to have a conditional, rather than a marginal, interpretation; as a result, we propose new techniques for estimation that are built on this distinction. We emphasize the use of these methods for sparse data due to the benefits of borrowing strength across subjects in this setting, but note that the same considerations exist in the dense case. Our approach is to make use of the generalized additive mixed model framework that inherently underlies model (2); we estimate quantities of interest either in a frequentist framework via a two-step procedure, or in Bayesian framework via a joint modeling algorithm. We show through numerical studies that accounting for the dependence in this way leads to improved mean estimation and to improved prediction of latent