Computational Statistics and Data Analysis 122 (2018) 101–114

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A joint design for functional data with application to scheduling ultrasound scans[☆]

So Young Park[a], Luo Xiao[b,*], Jayson D. Willbur[c], Ana-Maria Staicu[b], N. L'ntshotsholé Jumbe[d]

[a] Eli Lilly and Company, Indianapolis, IN, USA
[b] North Carolina State University, Raleigh, NC, USA
[c] Metrum Research Group LLC, Tariffville, CT, USA
[d] Bill & Melinda Gates Foundation, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 9 May 2017
Received in revised form 12 January 2018
Accepted 13 January 2018
Available online 31 January 2018

Keywords:
Covariance function
Functional data analysis
Fetal growth
Longitudinal data
Prediction

ABSTRACT

A joint design for sampling functional data is proposed to achieve optimal prediction of both functional data and a scalar outcome. The motivating application is fetal growth, where the objective is to determine the optimal times to collect ultrasound measurements in order to recover fetal growth trajectories and to predict child birth outcomes. The joint design is formulated using an optimization criterion and implemented in a pilot study. Performance of the proposed design is evaluated via simulation study and application to fetal ultrasound data.

1. Introduction

Functional data analysis has been a popular statistical research area for the last two decades and has found application in many fields such as brain imaging (Jiang et al., 2009; Greven et al., 2010; Reiss and Ogden, 2010; Lindquist, 2012; Lu and Marron, 2014; Park and Staicu, 2015), biosignals (Crainiceanu et al., 2012; Randolph et al., 2012; Goldsmith and Kitago, 2016), genetics (Tang and Müller, 2009; Reimherr and Nicolae, 2014) and wearable computing (Morris et al., 2006; Li et al., 2014; Xiao et al., 2015). For a comprehensive treatment of functional data analysis see Ramsay and Silverman (2002, 2005) and Horváth and Kokoszka (2012).

This paper considers sampling design for noisy growth data. The motivation arises from the study of fetal growth, where measurements of fetal size may be obtained during pregnancy using ultrasound. And the particular question to be addressed is: when a fixed number of ultrasound scans will be taken during pregnancy, what are the optimal time points for data collection? Optimality can be defined either in terms of recovering individual fetal growth trajectories or in terms of predicting a birth outcome, such as birth weight. However, in practice it may be important to predict both individual growth trajectories and birth outcomes, and in such cases a joint optimality criterion must be formulated. We also consider the closely related question of the number of ultrasound scans required to achieve a desired level of optimality.

We address this question within the functional data framework. Design for functional data has received some interest recently. For example, Ferraty et al. (2010) considered a nonparametric model with a scalar response and a functional

[☆] Supplementary materials are available with this article at the *Computational Statistics and Data Analysis* website.

* Correspondence to: Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.
E-mail address: lxiao5@ncsu.edu (L. Xiao).

https://doi.org/10.1016/j.csda.2018.01.009

predictor and Delaigle et al. (2012) studied a similar problem for classifying and clustering functional data. Both methods are restricted to densely sampled functional data and focus on dimensionality reduction for a dense functional predictor. And for spatially correlated functional data, Rasekhi et al. (2014) and Bohorquez et al. (2015) considered the problem of selecting spatial sampling points.

Design for functional data has also been extended to longitudinal data. Ji and Müller (2017) proposed prediction-based criteria for sampling functional data with the target of either recovering individual functions or predicting a scalar outcome. Wu et al. (2017) exploited the mixed effects model representation of functional data and proposed a design criterion based on Fisher's information matrix of eigenvalues of the covariance function. There are several limitations with these approaches. Wu et al. (2017) focused on recovering individual functions, while Ji and Müller (2017) were limited to the study of design separately and did not consider a joint design, which is the focus of our data application. In addition, in these works the number of design points was pre-fixed and no data-driven method was developed. Finally, Ji and Müller (2017) did not compare functional data models versus parametric mixed effects models for prediction-based designs. Our work addresses these gaps.

Following early work on design, such as Ylvisaker (1987) and the references therein and recent work by Ji and Müller (2017), we consider prediction-based designs and propose a unified design criterion for both recovering individual functions as well as predicting scalar outcomes from a functional predictor. We also propose a practical data-driven method for selecting the number of design points, building on the result that the larger the number of design points, the better the prediction will be (see Theorem 1). Finally we conduct a comprehensive simulation study to evaluate the performance of functional data models as compared to parametric mixed effects models, and demonstrate numerically that functional data models might be preferred over parametric mixed effects models for prediction-based optimal designs for longitudinal data.

The rest of the paper is organized as follows. In Section 2 we introduce functional data models and propose a unified prediction-based design criterion for sampling functional data. In Section 3 we study the theoretic properties of the proposed design. In Section 4 we discuss implementation of the design and propose a data-driven method for selecting the number of design points. In Section 5 we illustrate the proposed method using a fetal ultrasound data. In Section 6, we investigate the performance of the design via simulation studies.

2. Optimal design for functional data

In this section, we first describe functional data models and then formulate two optimal design problems for sampling functional data: one design targets accurate prediction of individual functions while the other targets accurate prediction of a scalar outcome. Then, we propose a unified design criterion that targets both recovering individual functions and predicting a scalar outcome. In particular, the unified design contains the previous two designs as special cases.

2.1. Statistical models

Consider a random function $X(t)(t \in \mathcal{T})$ defined over a continuous and compact time domain $\mathcal{T}$. Suppose that $X(\cdot)$ is a Gaussian process with mean function $\mu(t) = \mathrm{E}\{X(t)\}$ and covariance function $r(s,t) = \mathrm{Cov}\{X(s), X(t)\}$. We assume that $X(\cdot)$ is square integrable in $\mathcal{T}$ and without loss of generality we let $\mathcal{T} = [0, 1]$.

In practice, $X(\cdot)$ is observed at a finite number of time points and contaminated with noise. Hence, for a random function $X_i(\cdot)$ with a subject index i observed at p time points $(t_1, \ldots, t_p)' \in \mathcal{T}^p$, the observations are

$$W_{ij} = X_i(t_j) + \epsilon_{ij}, \quad 1 \le j \le p, \tag{1}$$

where the ϵ_{ij} are i.i.d. $\mathcal{N}(0, \sigma_\epsilon^2)$ and independent of $X_i(\cdot)$.

Let Y be a scalar outcome with a functional predictor $X(\cdot)$. And consider the functional linear model

$$Y = \alpha + \int_{\mathcal{T}} \bar{X}(t)\beta(t)\mathrm{d}t + e, \tag{2}$$

where α is an intercept, $\bar{X}(t) = X(t) - \mu(t)$, $\beta(t)$ is a smooth coefficient function, and e is white noise independent of $X(\cdot)$ with mean zero and variance σ_e^2.

The fundamental element in functional data analysis is the covariance function $r(s,t)$. By Mercer's theorem, $r(s,t)$ can be written as $\sum_{l=1}^{\infty} \lambda_\ell \phi_\ell(s)\phi_\ell(t)$, where $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ is the collection of eigenvalues and the $\phi_\ell(\cdot)$ are the associated eigenfunctions which satisfy $\int_{\mathcal{T}} \phi_\ell(t)\phi_{\ell'}(t)\mathrm{d}t = 1_{\{\ell=\ell'\}}$. Here $1_{\{\cdot\}}$ is 1 if the condition inside the bracket holds and 0 otherwise. To ensure that $\beta(t)$ is identifiable, we assume that the coefficient function $\beta(t)$ can be written as $\sum_{\ell=1}^{K} \beta_\ell \phi_\ell(t)$, where the β_ℓ are scalars and, a possibly infinite K represents the number of non-zero eigenvalues.

2.2. Optimal design for predicting functions

Fix $p \ge 1$ and assume that p observations will be collected from a new subject. The goal is to select the p *optimal* sampling points in $\mathcal{T}$ for predicting the new subject's curve with the smallest possible error.