Statistical Methodology 18 (2014) 79–100

Contents lists available at ScienceDirect

Statistical Methodology

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

Bayesian covariance estimation and inference in latent Gaussian process models

Cecilia Earls *, Giles Hooker

Cornell University, United States

ARTICLE INFO

Article history:
Received 17 May 2013
Received in revised form
2 September 2013
Accepted 4 October 2013

Keywords:
Bayesian modeling
Covariance
Functional data
Functional regression
Smoothing

ABSTRACT

This paper describes inference methods for functional data under the assumption that the functional data of interest are smooth latent functions, characterized by a Gaussian process, which have been observed with noise over a finite set of time points. The methods we propose are completely specified in a Bayesian environment that allows for all inferences to be performed through a simple Gibbs sampler. Our main focus is in estimating and describing uncertainty in the covariance function. However, these models also encompass functional data estimation, functional regression where the predictors are latent functions, and an automatic approach to smoothing parameter selection. Furthermore, these models require minimal assumptions on the data structure as the time points for observations do not need to be equally spaced, the number and placement of observations are allowed to vary among functions, and special treatment is not required when the number of functional observations is less than the dimensionality of those observations. We illustrate the effectiveness of these models in estimating latent functional data, capturing variation in the functional covariance estimate, and in selecting appropriate smoothing parameters in both a simulation study and a regression analysis of medfly fertility data.

1. Introduction

This paper describes a complete Bayesian methodology for estimating parameters in a Gaussian process model from partially observed data in the context of functional data analysis (FDA). FDA is

* Corresponding author. Tel.: +1 607 277 2571.
E-mail address: cae79@cornell.edu (C. Earls).

http://dx.doi.org/10.1016/j.stamet.2013.10.001

concerned with the analysis of replicated smooth random processes over a continuous domain, most commonly time which we write as $X_1(t), \ldots, X_N(t)$. See Ramsay and Silverman [14] for an overview of models and examples. Many of these methods can be thought of as extensions of multivariate analysis to infinite-dimensional data, combined with smoothing methods to ensure the stability of estimates. While it is unrealistic to assume that the processes in question are observed exactly at all times, much of the early work in FDA assumed that observations are precise and frequent enough that pre-smoothing could be employed to obtain representations of the smooth processes.

In this context, recent attention has been given to situations in which each replicate process is observed noisily, infrequently, and possibly at irregularly spaced intervals yielding a model of the form $Y_i(t_{ij}) = X_i(t_{ij}) + \epsilon_{ij}$; see Yao et al. [18] for pioneering work. This framework essentially makes each postulated process an infinite dimensional latent variable. Our contribution to this work is to describe a complete Bayesian framework that models the underlying latent functions as being generated by a Gaussian Process (GP) model. We establish priors and sampling methods for the unknown mean function and covariance surface in this model and show that, as with much of the previous work in FDA, these can be developed out of natural extensions of Bayesian approaches to multivariate data where additional smoothness constraints can be incorporated into priors for both the mean and covariance parameters. These factors make this a natural framework for functional data analysis and allow efficient Monte Carlo estimation via a Gibbs sampler. An important part of this paper is to demonstrate that we can effectively use finite-dimensional approximations to the posterior distribution to provide computationally efficient access to posterior inference for these parameters.

This paper extends the current literature on functional data analysis by providing a complete Bayesian framework for inference in FDA that includes non-parametric modeling and inference for functional parameters in a single estimation process. This then allows variance due to the estimation of mean and variance parameters to be incorporated within inferential procedures and provides a framework for inference in more complex models in which latent functional processes need not be directly observed at all. A particularly important aspect of this is in the estimation of a covariance surface within our methods. In related work, Yao et al. [18] proposed smoothing a method-of-moments representation of the covariance surface obtained at pairs of observation time points and reconstructing the latent functions via a principal components analysis (PCA) of this surface. This approach was also followed in Goldsmith et al. [7] in the context of regressing a response on the estimated scores. Crainiceanu and Goldsmith [4] presented a Bayesian version of this regression in which the uncertainty in the latent PCA scores is accounted for, but relied on a pre-estimate of the covariance surface via methods similar to Yao et al. [18]; none of these methods incorporate uncertainty in the covariance estimate into inferential procedures. Covariance has also been estimated via a spline representation with a penalized log-likelihood, Kauermann and Wegener [8] and also Cai and Yuan [2]. Within Bayesian methods, Linde [11] considers the covariance of a set of spline coefficients to represent functional data, and Kaufman and Sain [9] employ a class of covariance functions that are characterized by a small number of parameters. Nguyen and Gelfand [13] take a Bayesian approach to classifying functions that similar to our models are noisily observed; two major areas in which our models are different are (1) Nguyen and Gelfand use canonical components to model the latent functions and (2) they model their covariance functions parametrically which implicitly assumes there are no long-range dependencies in the functions.

Our approach differs from these in both incorporating all parameters in a GP model within a single hierarchical Bayesian analysis and in removing restrictions on the class of covariance surfaces that are used. We demonstrate that smoothness assumptions usually made directly on the $X_i(t)$ can be effectively reproduced within priors on the mean function and covariance surface. We also include priors on smoothing parameters, avoiding the need for cross validation and show that this has the effect of providing additional numerical stability to our Gibbs sampling procedure. Behseta et al. [1] proposed a hierarchical model to describe variation in the covariance function, but required a separate smoothing procedure from which plug-in estimators are derived.

A crucial component of this paper is the demonstration that posterior information for our model can be reliably obtained by applying our Gibbs sampler to the evaluation of all parameters at a common set of points followed by linear or bi-linear interpolation between them. This effectively reduces the problem to one of Bayesian estimation in a multivariate latent-vector model. In GP