Computational Statistics and Data Analysis 70 (2014) 362–372

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimator selection and combination in scalar-on-function regression

Jeff Goldsmith [a,*], Fabian Scheipl [b]

[a] Department of Biostatistics, Columbia Mailman School of Public Health, United States
[b] Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany

ARTICLE INFO

Article history:
Received 22 March 2013
Received in revised form 6 October 2013
Accepted 8 October 2013
Available online 16 October 2013

Keywords:
Cross validation
Functional linear model
Model stacking
Super learning

ABSTRACT

Scalar-on-function regression problems with continuous outcomes arise naturally in many settings, and a wealth of estimation methods now exist. Despite the clear differences in regression model assumptions, tuning parameter selection, and the incorporation of functional structure, it remains common to apply a single method to any dataset of interest. In this paper we develop tools for estimator selection and combination in the context of continuous scalar-on-function regression based on minimizing the cross-validated prediction error of the final estimator. A broad collection of functional and high-dimensional regression methods is used as a library of candidate estimators. We find that the performance of any single method relative to others can vary dramatically across datasets, but that the proposed cross-validation procedure is consistently among the top performers. Four real-data analyses using publicly available benchmark datasets are presented; code implementing these analyses and facilitating the application of proposed methods on future datasets is available in a web supplement.

1. Introduction

The problem of predicting continuous scalar outcomes from functional predictors has received high levels of interest in recent years, driven in part by a proliferation of complex datasets and by an increase in computational power. Although there are now many approaches to this problem, including several techniques for the popular functional linear model and methods for a number of data-generating scenarios, it is rare for a practitioner to apply more than one scalar-on-function regression method to any dataset. Doing so would potentially yield improved predictions of outcomes or new insights into scientific processes; indeed, the choice of regression model and estimation technique (a process we will refer to as estimator selection) is important and can dramatically affect prediction of outcomes and interpretation of results.

In this manuscript we develop approaches to facilitate the comparison and combination of many scalar-on-function estimation methods. We first focus on estimator selection, or the choice of a single estimator from a large collection of candidates, and then on the dynamic combination of approaches to yield an optimal ensemble estimator of the association between a scalar outcome and a functional predictor. Our proposed approaches are based on estimator selection through minimizing cross-validated loss (Breiman, 1996; Dudoit and van der Laan, 2005; van der Laan and Dudoit, 2003; Wolpert, 1992), referred to variously in the literature as model stacking and super learning. We adapt these strategies to the setting in which predictors are both high dimensional and spatially structured. Publicly available software allows easy comparison and selection of methods for predicting scalar outcomes from functional predictors.

* Corresponding author. Tel.: +1 212 342 4599.
E-mail address: jeff.goldsmith@columbia.edu (J. Goldsmith).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.10.009

Many approaches to scalar-on-function regression for continuous outcomes are now available. In the context of the functional linear model (described below), techniques include functional principal components regression (Ramsay and Silverman, 2005) and partial least squares (Reiss and Ogden, 2007), and penalized spline methods (Cardot et al., 2003; Goldsmith et al., 2011a; Marx and Eilers, 1999; James et al., 2009). Extensions of the functional linear model include the functional generalized additive model (McLean et al., in press), the functional additive model (James and Silverman, 2005), and single-index regressions (Eilers et al., 2009). A point-impact model was proposed in Lindquist and McKeague (2009) and a Bayesian hierarchical regression kernel method was developed in Woodard et al. (in press). In addition, high-dimensional regression and machine learning methods that are not specifically designed for structured functional data can nonetheless be applied to such datasets. Such methods include ridge regression, lasso and elastic net (Friedman et al., 2010), classification and regression trees (Breiman et al., 1984), boosting (Freund and Schapire, 1995), random forests (Breiman, 2001), and support vector machines (Suykens and Vandewalle, 1999). These methods are not necessarily designed for functional data, but could be applied to scores resulting from a truncated functional principal component analysis or from other reduced rank basis representations to give hybrid functional methods. Many of the approaches mentioned above are accompanied by software implementations.

We are motivated by a desire to optimally predict continuous scalar outcomes from functional data, acknowledging that no one method will be universally superior, and therefore pursue estimator selection and ensembling to compare and combine competing methods. To demonstrate the practical significance of this approach, we consider four real-data examples in this manuscript. First we consider the standard Canadian weather dataset, in which daily temperature measurements are used to estimate log annual precipitation at 35 monitoring stations. Next we analyze the Tecator dataset, which consists of 215 near-infrared (NIR) absorbance spectra of meat samples used as predictors of fat content of the sample. Third we analyze a diffusion tensor imaging (DTI) dataset, where the goal is predicting a scalar measure of cognitive function from functional summaries of intracranial white matter microstructure using 334 observations. Finally we examine an additional NIR spectra dataset, which consists of 72 samples of cookie dough in which the sucrose content is of interest. These examples illustrate several practical issues related to the application of scalar-on-function regression methods, including the differential performance of individual methods across datasets, the value of applying, selecting, and ensembling multiple methods, and the computational concerns in the proposed techniques. All datasets considered are publicly available, and the code implementing each analysis is available as a web supplement.

The remainder of the manuscript is organized as follows. A broad selection of approaches for functional regression is discussed in Section 2, while estimator selection and ensembling are detailed in Section 3. Real data analyses are presented in Section 4. We close with a discussion in Section 5.

2. Existing methods for continuous scalar-on-function regression

We observe data $[Y_i, W_i(s)]$ for subjects $1 \leq i \leq I$ where Y_i is a continuous outcome and $W_i(s)$, without loss of generality assuming $s \in [0, 1]$, is the functional predictor of interest. In practice the curves $W_i(s)$ are observed on a discrete grid $\{s_{ij}\}_{j=1}^{J_i}$ that is potentially sparse and subject-specific, and often observations are subject to measurement error. Preprocessing steps such as smoothing or functional principal components analysis (FPCA) can be used to reduce the effect of measurement error and obtain curves on a dense common grid $\{s_j\}_{j=1}^{J}$; in this exposition we will assume data in this form. This section reviews existing methods for estimating the regression function $\psi_0(W(s)) = \mathrm{E}[Y \mid W(s)]$. While we attempt to be thorough, this review is not exhaustive. Our discussion focuses only on a single functional predictor although several of the approaches discussed allow multiple functional predictors or the inclusion of non-functional covariates, both of which are important in practice.

2.1. Functional linear model

The functional linear model (FLM) extends the standard multiple linear regression model to functional predictors. Thus we assume

$$Y_i = \int_0^1 W_i(s)\beta(s)ds + \epsilon_i \tag{1}$$

where $\epsilon_i \sim \mathrm{N}\left[0, \sigma^2\right]$ and $\beta(s)$ is the coefficient function. The FLM seeks to minimize the sum of squared errors $\|Y - \int_0^1 W(s)\beta(s)ds\|^2$ where $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}^T\boldsymbol{v}}$. It is additionally assumed, either implicitly or explicitly, that the coefficient function $\beta(s)$ is smooth in some sense over its domain; such an assumption respects the local structure inherent in the predictor and avoids the problem of an ill-posed regression when $J \geq I$. The functional linear model is perhaps the most common approach for scalar-on-function regression, and many techniques have been proposed to estimate the coefficient function $\beta(s)$ based on different assumed forms of this function. The coefficient function $\beta(s)$ is an interpretable object: locations with large $|\beta(s)|$ are influential for the outcome, and the direction of the association is given by the sign of the coefficient function.

Functional principal components regression (FPCR) is based on an FPCA decomposition of the functional predictors (Ramsay and Silverman, 2005). Specifically, curves are approximated using $W_i(s) \approx \sum_{k=1}^{K_W} c_{ik}\phi_k(s)$ where $\boldsymbol{c}_i = \{c_{ik}\}_{k=1}^{K_W}$