Journal of Multivariate Analysis 140 (2015) 395–402

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Minimax prediction for functional linear regression with functional responses in reproducing kernel Hilbert spaces

Heng Lian

School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, 2052, Australia

ARTICLE INFO

Article history:
Received 2 February 2014
Available online 16 June 2015

AMS subject classifications:
62G05
62G20

Keywords:
Functional data
Functional response
Minimax convergence rate
Regularization

ABSTRACT

In this article, we consider convergence rates in functional linear regression with functional responses, where the linear coefficient lies in a reproducing kernel Hilbert space (RKHS). Without assuming that the reproducing kernel and the covariate covariance kernel are aligned, convergence rates in prediction risk are established. The corresponding lower bound in rates is derived by reducing to the scalar response case. Simulation studies and two benchmark datasets are used to illustrate that the proposed approach can significantly outperform the functional PCA approach in prediction.

1. Introduction

The literature contains an impressive range of functional analysis tools for various problems including exploratory functional principal component analysis, canonical correlation analysis, classification and regression. Two major approaches exist. The more traditional approach, masterfully documented in the monograph [16], typically starts by representing functional data by an expansion with respect to a certain basis, and subsequent inferences are carried out on the coefficients. The most commonly utilized basis includes B-spline basis for nonperiodic data and Fourier basis for periodic data. Another line of work by the French school [10], taking a nonparametric point of view, extends the traditional nonparametric techniques, most notably the kernel estimate, to the functional case. Some recent advances in the area of functional regression include Cardot et al. [6], Cai and Hall [4], Preda [15], Lian [13], Ait-Saidi et al. [2], Yao et al. [18], Crambes et al. [8], Ferraty et al. [11], Lian [14].

In this paper we study the functional linear regression problem of the form

$$Y(t) = \mu(t) + \int_0^1 \beta(t,s)X(s)\,ds + \epsilon(t), \tag{1}$$

where $Y, X, \epsilon \in L_2[0,1]$ and $E[\epsilon|X] = 0$, the same problem that appeared in [16,18,3,1,9]. In terms of methodology, the plan of attack we will give for (1) is most closely related to that of Crambes and Mas [9]. In this introduction, we will explain the methodology used in that paper and then the different assumptions we will make on $\beta(t,s)$.

Without loss of much generality, throughout the paper we assume $E(X) = 0$ and the intercept $\mu(t) = 0$, since the intercept can be easily estimated. The covariance operator of X is the linear operator $\Gamma = E(X \otimes X)$ where for $x, y \in L_2[0,1]$, $x \otimes y : L_2[0,1] \to L_2[0,1]$ is defined by $(x \otimes y)(g) = \langle y, g\rangle x$ for any $g \in L_2[0,1]$. Γ can also be represented by the bivariate

E-mail address: heng.lian@unsw.edu.au.

http://dx.doi.org/10.1016/j.jmva.2015.06.005
0047-259X/

function $\Gamma(s,t) = E[X(s)X(t)]$. Using the same letter Γ to denote both the operator and the bivariate function will not cause confusion in our context. We assume throughout the paper that $E\|X\|^4 < \infty$ which implies Γ is a compact operator. Then by the Karhunen–Loève Theorem there exists a spectral expansion for Γ,

$$\Gamma = \sum_{j=1}^{\infty} \lambda_j \varphi_j \otimes \varphi_j,$$

where $\lambda_j \geq 0$ are the eigenvalues with $\lambda_j \to 0$ and $\{\varphi_j\}$ are the orthonormalized eigenfunctions. Correspondingly, we have the representation $X = \sum_{j\geq 1} \gamma_j \varphi_j$ with $\gamma_j = \int X\varphi_j$. The random coefficient γ_j satisfies $E\gamma_j\gamma_k = \lambda_j I\{j = k\}$ where $I\{.\}$ is the indicator function.

By expanding β using the set of eigenfunctions, we write $\beta(t,s) = \sum_{j\geq 1} b_j(t)\varphi_j(s)$ and (1) can be equivalently written as

$$Y(t) = \sum_{j\geq 1} b_j(t)\gamma_j + \epsilon(t).$$

Multiplying both sides above by γ_j and taking expectations, we obtain $b_j(t) = E[Y(t)\gamma_j]/\lambda_j$. Given i.i.d. data (X_i, Y_i), $i = 1, \ldots, n$, $\{\lambda_j, \varphi_j\}$ can be easily estimated by $\hat{\lambda}_j$ and $\hat{\varphi}_j$ obtained from the spectral decomposition of the empirical covariance operator and $E[Y(t)\gamma_j]$ can be approximated by the corresponding sample average. Thus the estimator proposed in Crambes and Mas [9] is

$$\hat{\beta}(t,s) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{\int X_i\hat{\varphi}_j}{\hat{\lambda}_j} Y_i(t)\hat{\varphi}_j(s).$$

Note that the infinite sum over j has been truncated as some point k for regularization. One intriguing point is that there is no regularization on $Y_i(t)$ necessary, in contrast with Yao et al. [18] where Y is observed sparsely with additional noise. This can also be seen from that $b_j(t)$ is not a priori constrained in any way. The reason is that only regularization of the covariance operator, which does not depend on Y, is necessary to avoid overfitting.

Minimax convergence rates of $E\|\int \hat{\beta}(t,s)X(s)ds - \int \beta(t,s)X(s)ds\|^2$ were shown in Crambes and Mas [9]. Although no assumption is needed on b_j except that the regression operator is Hilbert–Schmidt, slowly decaying $\|b_j\|$ as j increases will lead to slow convergence rate of the regression function. Given that $\|b_j\|$'s are the coefficients of $\beta(t,s)$ in terms of the basis φ_j, which is a characteristic of the predictor, there is no a priori reason why this basis should provide a good representation of β in the sense that $\|b_j\|$ will decay fast. A more general assumption for β may be on its smoothness, which makes a reproducing kernel Hilbert space (RKHS) approach an interesting alternative. Such arguments have led to the developments in [19,5] for the scalar response models. While Crambes and Mas [9] is based on Cardot et al. [7] for scalar response models, ours is based on Cai and Yuan [5].

The rest of the article is organized as follows. In Section 2, we propose an estimator for β with an RKHS approach where the reproducing kernel and the covariance kernel are not necessarily aligned. Following Cai and Yuan [5], we say that the two kernels are perfectly aligned if they have the same sequence of eigenvalues when ordered according to the eigenvalues. We establish the minimax rate of convergence in prediction risk by deriving both the upper bound and the lower bound. In Section 3, we present some simulation studies to show that the RKHS approach could significantly outperform the functional PCA approach when the kernels are mis-aligned. This advantage is further illustrated on two benchmark datasets which show better prediction performance using our approach. We conclude in Section 4 with some discussions. The technical proofs are relegated to Appendix.

Finally, we list some notations and properties regarding different norms to be used. For any operator $\mathcal{F}$, we use $\mathcal{F}^{\mathrm{T}}$ to denote its adjoint operator. If $\mathcal{F}$ is self-adjoint and nonnegative definite, $\mathcal{F}^{1/2}$ is its square-root satisfying $\mathcal{F}^{1/2}\mathcal{F}^{1/2} = \mathcal{F}$. For $f \in L_2$, $\|f\|$ denotes its L_2 norm. For any operator $\mathcal{F}$, $\|\mathcal{F}\|_{op}$ is the operator norm $\|\mathcal{F}\|_{op} := \sup_{\|f\|\leq 1}\|\mathcal{F}f\|$. The trace norm of an operator $\mathcal{F}$ is $\mathrm{Trace}(\mathcal{F}) = \sum_k \langle(\mathcal{F}^{\mathrm{T}}\mathcal{F})^{1/2}e_k, e_k\rangle$ for any orthonormal basis $\{e_k\}$ of L_2. $\mathcal{F}$ is a trace class operator if its trace norm is finite. The Hilbert–Schmidt norm of an operator is $\|\mathcal{F}\|_{HS} = (\sum_{j,k}\langle\mathcal{F}e_j, e_k\rangle^2)^{1/2} = (\sum_j \|\mathcal{F}e_j\|^2)^{1/2}$. An operator is a Hilbert–Schmidt operator if its Hilbert–Schmidt norm is finite. From the definition it is easy to see that $\mathrm{Trace}(\mathcal{F}^{\mathrm{T}}\mathcal{F}) = \mathrm{Trace}(\mathcal{F}\mathcal{F}^{\mathrm{T}}) = \|\mathcal{F}\|_{HS}^2$. Furthermore, if $\mathcal{F}$ is a Hilbert–Schmidt operator and $\mathcal{G}$ is a bounded operator, then $\mathcal{F}\mathcal{G}$ is also a Hilbert–Schmidt operator with $\|\mathcal{F}\mathcal{G}\|_{HS} \leq \|\mathcal{F}\|_{HS}\|\mathcal{G}\|_{op}$.

2. Methodology and convergence rates

Following Wahba [17], a RKHS H is a Hilbert space of real-valued functions defined on, say, the interval $[0, 1]$, in which the point evaluation operator $L_t : H \to R, L_t(f) = f(t)$ is continuous. By Riesz representation theorem, this definition implies the existence of a bivariate function $K(s,t)$ such that

$K(s,\cdot) \in H$, for all $s \in [0,1]$

and (reproducing property)

for every $f \in H$ and $t \in [0,1]$, $\langle K(t,\cdot), f\rangle_H = f(t)$.