Journal of Multivariate Analysis 124 (2014) 150–165

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Series expansion for functional sufficient dimension reduction

Heng Lian [a,*], Gaorong Li [b]

[a] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371, Singapore

[b] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China

ARTICLE INFO

Article history:
Received 15 June 2012
Available online 5 November 2013

AMS subject classification:
62H12

Keywords:
Functional principal component analysis
Polynomial splines
Sliced average variance estimation
Sliced inverse regression

ABSTRACT

Functional data are infinite-dimensional statistical objects which pose significant challenges to both theorists and practitioners. Both parametric and nonparametric regressions have received attention in the functional data analysis literature. However, the former imposes stringent constraints while the latter suffers from logarithmic convergence rates. In this article, we consider two popular sufficient dimension reduction methods in the context of functional data analysis, which, if desired, can be combined with low-dimensional nonparametric regression in a later step. In computation, predictor processes and index vectors are approximated in finite dimensional spaces using the series expansion approach. In theory, the basis used can be either fixed or estimated, which include both functional principal components and *B*-spline basis. Thus our study is more general than previous ones. Numerical results from simulations and a real data analysis are presented to illustrate the methods.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

There has recently been increased interest in the statistical modeling of functional data. In many experiments, functional data appear as the basic unit of observations. As a natural extension of the multivariate data analysis, functional data analysis provides valuable insights into these problems. Compared with the discrete multivariate analysis, functional analysis takes into account the smoothness of the high dimensional covariates, and often suggests new approaches to the problems that have not been discovered before. Even for nonfunctional data, the functional approach can often offer new perspectives on the old problem.

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 [29], 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 include *B*-spline basis for nonperiodic data and Fourier basis for periodic data. Another line of work by the French school [16], 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. [5]; Cai and Hall [4]; Aneiros-Perez and Vieu [3]; Preda [28]; Ait-Saidi et al. [2]; Aguilera et al. [1]; Wong et al. [30]; Yao et al. [32]; Ait-Saidi et al. [2]; Crambes et al. [13].

As an extension of classical linear regression, parametric functional linear regression has achieved exclaimed success in many real problems, although it can be argued that the structural constraint is too stringent. On the other hand,

* Corresponding author.
E-mail address: henglian@ntu.edu.sg (H. Lian).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.10.019

nonparametric functional regression is more flexible but typically suffers from poor convergence rate [17]. To address these problems, Chen et al. [6] studied functional single-index and multiple-index models.

Here we consider an alternative semiparametric approach based on sufficient dimension reduction. In functional context, we assume

$$Y = g(\langle\beta_1, X\rangle, \ldots, \langle\beta_K, X\rangle, \epsilon), \tag{1}$$

where $\langle , \rangle$ is the usual inner product in $L_2[0, 1]$. Thus the response Y only depends on the predictor through K indices obtained by projecting onto K directions. Since g is unknown, the K directions, referred to as dimension reduction directions, are not identifiable. In the multivariate case, the space spanned by them (referred to as a dimension reduction subspace, or drs) is identifiable under mild assumptions, however such assumptions are not known in the functional context yet. Thus we will not use the concept of the central space which is popularly used in the dimension reduction literature [7,34]. The reason is that for functional data there is no corresponding theory for the existence and uniqueness of the central space. This may be due to that density for functional data is a tricky concept to work with. In the literature of functional SIR, researchers typically work with a drs, even though there might be multiple drs's. Although a unique drs is generally not identifiable, useful methodology is still possible. The approach of dimension reduction is particularly useful in an exploratory stage of statistical analysis since very few structural assumptions are imposed in (1). In particular, it is not necessary to assume the different indices act additively as usually assumed in multiple-index models, and the error also is not necessarily additive on mean, or homogeneous. After the dimension reduction directions are found, in particular if there are only a small number of significant directions, one can use traditional nonparametric approaches to study the relationships between responses and the few indices. This second stage typically involves additional structural assumptions such as additive errors.

There exist quite a few different methods aimed at estimating the dimension reduction space [24,12,25,37,38]. Among these sliced inverse regression (SIR) and sliced average variance estimation (SAVE) are probably the most popular. Both required linearity assumption of the predictors. However, SIR will fail when $E[X|Y] = 0$ which motivated the use of SAVE. On the other hand, SAVE requires an additional assumption on the distribution of predictors.

Adapting SIR to functional context has been proposed in [18] based on functional principal component analysis on the random predictor process. In particular the predictor process is approximated by a truncation of the Karhunen–Loève expansion, using the eigenfunctions as the basis. The basic procedure is to (i) approximate the functional predictors with series expansion using certain basis and obtain the coefficients; (ii) perform dimension reduction using the finite-dimensional coefficients as the predictors; (iii) use directions obtained in (ii) as the coefficients of the basis to finally obtain the direction in functional space. It turns out this computational procedure is correct only when the basis is orthonormal, and we will detail the general algorithm in Section 4.

In terms of theory, Ferré and Yao [18] assumes that the number of slices is fixed which works well for discrete responses, but is only an approximation for continuous responses. On the other hand, the kernel estimate used in [19] was later shown to require much stronger assumptions [9].

Our contributions in this study are summarized as follows. First, our theory for SIR allows various basis systems, either fixed in advanced or estimated from data. Second, our theory works for both categorical and continuous response Y. Third and most importantly, we extend SAVE to the functional context which has not been considered before.

2. SIR and SAVE

Let Y be a real random response and $X \in L_2[0, 1]$ the random functional predictor. In this article, we assume the entire trajectory of noise-free process X is observed. When the process is densely measured, this is a reasonable assumption. For simplicity, we assume $EX = 0$. We also assume the fourth moment of X exists, that is $E\|X\|^4 < \infty$. The (population) covariance operator of X is given by $\Gamma = E(X \otimes X)$, where for any $x, y \in L_2[0, 1]$, $x \otimes y$ denotes the linear operator $L_2[0, 1] \to L_2[0, 1]$ such that $(x \otimes y)(z) = \langle x, z\rangle y$. Using the well-known Karhunen–Loève expansion, we can write

$$X = \sum_{j=1}^{\infty} \xi_j \phi_j,$$

where $E\xi_j^2 = \lambda_j$ are the eigenvalues and ϕ_j are the eigenfunctions. We assume all the eigenvalues, $\lambda_1 > \lambda_2 > \cdots > 0$ are distinct and positive, as usually assumed in the functional data literature [22,18]. If some eigenvalues are zero, the components of β_k in the kernel space of Γ cannot be identified. We focus on the estimation of the space spanned by K linearly independent directions $\beta_1, \ldots, \beta_K$, which is called a dimension reduction subspace (drs) and denoted by $\mathcal{S}$. Let $\Gamma\mathcal{S}$ be the space spanned by $\Gamma\beta_1, \ldots, \Gamma\beta_K$.

Let $B_X = (\langle\beta_1, X\rangle, \ldots, \langle\beta_K, X\rangle)$. The principle of SIR and SAVE is based on the following result with proofs omitted, which is a direct extension of the multivariate case.

Theorem 1. (a) [18] *Suppose for all $b \in L_2[0, 1]$, the conditional expectation $E(\langle b, X\rangle|B_X)$ is linear in $\langle\beta_1, X\rangle, \ldots, \langle\beta_K, X\rangle$. Then $E(X|Y) \in \Gamma\mathcal{S}$. Obviously, if the linearity assumption is true for all drs's, then $E(X|Y) \in \Gamma(\cap_{\mathcal{S} \text{ is a drs}} \mathcal{S})$. Note that generally, $\cap_{\mathcal{S} \text{ is a drs}} \mathcal{S}$ is not guaranteed to be a drs, but s spanning system for the intersection may still be useful in practice.*

(b) *If in addition* $\mathrm{Var}(X|B_X)$ *is nonrandom,* $\Gamma - \mathrm{Var}(X|Y) \in \Gamma\mathcal{S}$. *Similarly, if this nonrandomness assumption holds for all drs's,* $\Gamma - \mathrm{Var}(X|Y) \in \Gamma(\cap_{\mathcal{S} \text{ is a drs}} \mathcal{S})$.

In the statement of the theorem above $\mathrm{Var}(X|B_X)$ denotes the conditional covariance operator of X, which can also be written as $E[(X - E(X|B_X)) \otimes (X - E(X|B_X))|B_X]$. With abuse of notation, $\Gamma - \mathrm{Var}(X|Y) \in \Gamma\mathcal{S}$ in fact means $(\Gamma - \mathrm{Var}(X|Y))\beta \in \Gamma\mathcal{S}$ for all $\beta \in L_2[0, 1]$. The two conditions in (a) and (b) above constrain the marginal distribution of the predictors, not the conditional distribution of $Y|X$ as is typical in regression. Both hold when X is a Gaussian process, although Gaussianity is not necessary.

Note that in the literature of sufficient dimension reduction in the Euclidean space, one usually simplifies the investigations somewhat by focusing on the standardized predictor $Z = \Gamma^{-1/2}X$. However, under the functional context, such standardization generally does not make sense. The reason is that by the Karhunen–Loève expansion, formally we have $\Gamma^{-1/2}X = \sum_{i=1}^{\infty}(\xi_i/\sqrt{\lambda_i})\phi_j$, but $E\xi_i^2/\lambda_i = 1$ implies $\sum_{i=1}^{\infty}(\xi_i/\sqrt{\lambda_i})\phi_j$ is not a well-defined element in $L_2[0, 1]$ with probability one.

Based on Theorem 1, functional SIR estimates a drs as the eigenspace of $\Gamma^{-1}\mathrm{Var}(E[X|Y])$. In the multivariate case, the SAVE estimator is defined by the eigenspace of $E[(I - \mathrm{Var}(Z|Y))^2]$ where $Z = (E[XX^T])^{-1/2}X$ is the standardized covariate. As noted above, such standardization is not possible for functional predictor, but we can define functional SAVE based on the same principle. We could use the eigenspace of $\Gamma^{-1}E[(\Gamma - \mathrm{Var}(X|Y))^2]$ to estimate a drs, but this does not reduce to SAVE in the multivariate case. Instead, we first note that $\Gamma - \mathrm{Var}(X|Y) \in \Gamma\mathcal{S}$ implies $\Gamma^{1/2} - \mathrm{Var}(X|Y)\Gamma^{-1/2} \in \Gamma\mathcal{S}$ which in turn implies $(\Gamma^{1/2} - \mathrm{Var}(X|Y)\Gamma^{-1/2})(\Gamma^{1/2} - \mathrm{Var}(X|Y)\Gamma^{-1/2})^* = \Gamma - 2\mathrm{Var}(X|Y) + \mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y) \in \Gamma\mathcal{S}$ where $(.)^*$ denotes the adjoint operator. Thus a drs can be obtained from the eigenspace of $\Gamma^{-1}E[(\Gamma - 2\mathrm{Var}(X|Y) + \mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y))]$. In general, the eigenspace of these operators constructed by SIR and SAVE is only a subspace of $\mathcal{S}$ and we can only hope to be able to recover this subspace. Nevertheless, for theoretical analysis, following the dimension reduction literature, we assume the eigenvectors in either the SIR or the SAVE approach exhaustively span $\mathcal{S}$. Furthermore, to ease notation, these eigenvectors are still denoted by $\beta_1, \ldots, \beta_K$ (even though eigenvectors for the two operators constructed in SIR and SAVE are obviously different). We assume the K eigenvalues of $\Gamma^{-1}\mathrm{Var}(E[X|Y])$ and $\Gamma^{-1}E[(\Gamma - 2\mathrm{Var}(X|Y) + \mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y))]$ are distinct for simplicity so that all eigenvectors associated with nonzero eigenvalues can be identified.

Since the domain of Γ^{-1} or $\Gamma^{-1/2}$ is not the whole $L_2[0, 1]$, these formal expressions might not be well-defined, which motivates the following proposition. Note that the sufficient conditions given in the theorem below do not depend on whether model (1) is true or not. If the model (1) is indeed true, then these expressions are automatically well-defined by Theorem 1.

Proposition 1. *If*

$$\sum_j \lambda_j^{-2} \sum_i (E[E(\xi_i|Y)E(\xi_j|Y)])^2 < \infty, \tag{2}$$

then $\Gamma^{-1}\mathrm{Var}(E(X|Y))$ *is well-defined. If*

$$E\left[\left(\sum_j \lambda_j^{-2} \sum_i \mathrm{Cov}_{ij|Y}^2\right)^2\right] < \infty, \tag{3}$$

then $\Gamma^{-1}E[\Gamma - 2\mathrm{Var}(X|Y) + \mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y)]$ *is well-defined, where* $\mathrm{Cov}_{ij|Y} = \mathrm{Cov}(\xi_i, \xi_j|Y)$.

3. Series expansion for functional data

We approximate both predictor X and direction β_k by a basis expansion

$$X(t) \approx \sum_{j=1}^{D} x_j B_j(t), \qquad \beta_k(t) \approx \sum_{j=1}^{D} b_{kj} B_j(t),$$

where D is the number of basis functions in approximating the functions. Various basis systems, such as Fourier bases, polynomial bases, and B-spline bases, can be used in this basis expansion. Bases estimated from data can also be used, with the most popular choice being the estimated eigenfunctions of Γ. Our methodology works for any basis system that satisfies some mild assumptions as detailed later. For specificity, we will emphasize the use of eigenfunctions estimated from functional PCA as well as B-spline basis. Theoretically, we can use a different number of basis functions in the approximation of different functions, or use a different basis system, but it is hard to choose multiple D's during estimation and it also makes notations more complicated to consider more than one basis system. Thus we only focus on one set of basis functions throughout the paper.

Since we allow the basis to be estimated from data, we distinguish between two D-dimensional subspaces of $L_2[0, 1]$, $\hat{S}_D$ and S_D, where the former is the space spanned by the estimated basis functions and the latter spanned by its population

counterpart. Let Π_D be the operator of projection onto the subspace S_D, which is finite-dimensional although the dimension diverges with sample size. Consider SIR first, where we want to estimate $\Gamma^{-1}\mathrm{Var}(E[X|Y])$. Even though the inverse Γ^{-1} exists, it is generally not a bounded operator. Thus we replace Γ by the finite-rank operator $\Gamma_D = \Pi_D\Gamma\Pi_D$ with (pseudo-)inverse $\Gamma_D^{-1} = \Pi_D\Gamma^{-1}\Pi_D$. Even though Γ^{-1} is not bounded, Γ_D^{-1} can be expected to be bounded. On the other hand, inverse of $\mathrm{Var}(E[X|Y])$ is not involved and thus it can be directly estimated without using finite-dimensional approximation, which we detail now.

Given an i.i.d. sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, define the order statistics $Y_{(1)} \le \cdots \le Y_{(n)}$, and let $X_{(i)}$ be the concomitant of $Y_{(i)}$. To obtain slicing estimator, the range of Y is divided into H slices. We assume each slice contains an equal number of observations, c, such that $n = Hc$ (n is assumed to be a multiple of c without loss of generality).

In the following we use $\|\cdot\|$ for multiple norms, with operator it indicates the operator norm, with matrix it indicates the spectral norm (maximum eigenvalue), with functions it indicates the L_2 norm, with vectors it indicates the l_2 (Euclidean) norm.

As a direct extension of sliced inverse regression to functional predictors, we estimate $\mathrm{Var}(E[X|Y])$ by

$$\widehat{\mathrm{Var}}(E[X|Y]) = \frac{1}{H}\sum_{h=1}^{H}\bar{X}_h \otimes \bar{X}_h,$$

where $\bar{X}_h$ is the sample average of the predictors in the hth slice. Below we demonstrate the convergence of $\widehat{\mathrm{Var}}(E[X|Y])$ to $\mathrm{Var}(E[X|Y])$. As seen in the proof, the convergence rate can be demonstrated in almost the same way as in the finite dimensional case. On the other hand, note that Hsing and Carroll [23]; Zhu and Ng [36] only considered another version of SIR based on estimation of $E(\mathrm{Var}(X|Y))$ and thus our proposition below is of interest even in the finite dimensional case, which seems to be missing in the literature in the case of $H \to \infty$. Note that Ferré and Yao [18] directly used $\|\widehat{\mathrm{Var}}(E[X|Y]) - \mathrm{Var}(E[X|Y])\| = O_p(n^{-1/2})$ without the following assumptions, thus their proof can only be applied to categorical responses, or as an approximation to continuous responses. Furthermore, for functional data, some modifications in the proof are necessary.

Denote $m(y) = E[X|Y = y]$ and let $\epsilon_i = X_i - m(Y_i)$. Here m is $L_2[0, 1]$-valued. Then $\epsilon_{(i)} = X_{(i)} - m(Y_{(i)})$ are conditionally independent given the order statistics $Y_{(i)}$ [31]. Smoothness condition on m is needed for the convergence of $\widehat{\mathrm{Var}}(E[X|Y])$. Let $P_n(T)$ be the set of all partitions $-T \le t_1 \le t_2 \le \cdots \le t_n \le T$ of the interval $[-T, T]$ where $T > 0$. We say a $L_2[0, 1]$-valued function $m(y)$ has total variation of order r if

$$\lim_{n\to\infty} n^{-r}\sup_{P_n(T)}\sum_{i=1}^{n-1}\|m(t_{i+1}) - m(t_i)\| = 0,$$

for any fixed $T > 0$. We say m is non-expansive on $(-\infty, -T] \cup [T, \infty)$ if there exists a nondecreasing function M such that for two points y_1 and y_2 both in $(-\infty, -T]$ or both in $[T, \infty)$, we have $\|m(y_1) - m(y_2)\| \le |M(y_1) - M(y_2)|$. These conditions are similar to those assumed in [23,36].

Proposition 2. *Assume that*

(i) $E\|X\|^4 < \infty$, *and all the eigenvalues of* Γ *are distinct and positive;*
(ii) *For some* $r > 0$, $m(y)$ *has a total variation of order* r, *and for some* $T_0 > 0$, m *is nonexpansive on* $(-\infty, -T_0] \cup [T_0, \infty)$, *with* M *satisfying* $\lim_{t\to\infty} M^4(t)P(|Y| > t) = 0$;
(iii) $c \sim n^\delta$ *for some* $0 < \delta < 1/2$.

Then $\|\widehat{\mathrm{Var}}(E[X|Y]) - \mathrm{Var}(E[X|Y])\| = O_P(n^{-\gamma})$ *with* $\gamma = \min\{\delta, 1 - \delta - r, 1/2 - \delta\}$.

Remark 1. As presented in the proposition above, we can only achieve the $n^{-1/4}$ convergence rate (with $\delta = 1/4$). However, the rate of order $O_p(n^{-1/2})$ is possible if we modify the assumptions. For example, if $m(y)$ is totally bounded on the entire range of Y, then we can use $r = 0$ in the proof where we have $\sum_i \|m(Y_{(i+1)}) - m(Y_{(i)})\| = O_p(1)$, and D_{122n}, D_{123n} defined in the proof will not appear. Then with $c \sim n^{1/2}$, it is obvious from the proof that the convergence rate is now $n^{-1/2}$. Alternatively, if Y is discrete taking a finite number of possible values, then it is easy to show that $\|\widehat{\mathrm{Var}}(E[X|Y]) - \mathrm{Var}(E[X|Y])\| = O_p(n^{-1/2})$ under mild assumptions ([18] directly used this without proof). When Y is discrete, the proof is easier than the proof of Proposition 2 and we do not present it here. Even when Y is continuous, we can construct a discrete version $\tilde{Y}$ of Y by quantization into H values. It is always true that $\mathcal{S}$ for $\tilde{Y}$ is a subset of $\mathcal{S}$ for Y, and when H is sufficiently large these two dimension reduction spaces are equal. This quantization approach is adopted in many asymptotic analysis of sliced inverse regression, including Li [24]; Duan and Li [15]; Cook and Ni [11].

Since Γ^{-1} is an unbounded operator and Γ^{-1} is involved in the estimation of a drs, it is necessary to regularize the operator, such as using the finite-rank approximation so that its generalized inverse is bounded. Corresponding to the population covariance operator, let $\widehat{\Gamma} = (\sum_{i=1}^n X_i \otimes X_i)/n$ be the natural moment estimator of Γ. Similarly, we define $\widehat{\Pi}_D$ as the projection onto $\widehat{S}_D$ and $\widehat{\Gamma}_D = \widehat{\Pi}_D\widehat{\Gamma}\widehat{\Pi}_D$. For functional PCA, $\widehat{S}_D$ is different from S_D since the eigenfunctions are estimated from spectral decomposition of $\widehat{\Gamma}$. On the other hand, when B-spline basis is used, $\widehat{S}_D = S_D$. Formally, with SIR approach, a drs is estimated as the space spanned by the top K eigenvectors of $\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}(E[X|Y])$, denoted by $\widehat{\beta}_1, \ldots, \widehat{\beta}_K$.

To state the theorem below, we define $s_D = \|\Gamma_D - \widehat{\Gamma}_D\|$, and let t_D be the smallest positive eigenvalue of Γ_D.

Theorem 2. *Suppose* $\|\mathrm{Var}(E[X|Y]) - \widehat{\mathrm{Var}}(E[X|Y])\| = O_p(n^{-\gamma})$ *for some* $0 < \gamma \le 1/2$. *If* $D \to \infty$, $s_D = o_p(t_D)$, $1/(\sqrt{n}t_D^2) \to 0$, *and* $1/(n^\gamma t_D^{3/2}) \to 0$, *then* $\|\widehat{\beta}_j - \beta_j\| = o_p(1), j = 1, \ldots, K$.

Remark 2. Compared to Ferré and Yao [18], the only additional assumption on D is that $1/(n^\gamma t_D^{3/2}) \to 0$. This is due to that we allow more general convergence rate for $\widehat{\mathrm{Var}}(E[X|Y])$. When $\gamma = 1/2$, this assumption is actually unnecessary here since it is implied by $1/(\sqrt{n}t_D^2) \to 0$.

Remark 3. Here we discuss the value of s_D and t_D defined above for the special cases of functional PCA and B-spline basis. In the former case, let $a_1 = 2\sqrt{2}/(\lambda_1 - \lambda_2)$ and $a_j = 2\sqrt{2}/\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$. Then following the arguments in [18], we have $\|\Gamma_D - \widehat{\Gamma}_D\| = O_p(\sum_{j=1}^D a_j/\sqrt{n})$. Furthermore, the minimum positive eigenvalue of Γ_D is λ_D and thus the condition that $s_D = o_p(t_D)$ reduces to $\sum_{j=1}^D a_j/(\sqrt{n}\lambda_D) \to 0$. For B-spline basis, we have obviously $\|\Gamma_D - \widehat{\Gamma}_D\| \le \|\Gamma - \widehat{\Gamma}\| = O_p(1/\sqrt{n})$. Thus we require that t_D approaches zero slower than $1/\sqrt{n}$.

Now we consider functional SAVE. The sample version of the condition variance of X given Y in each slice is $\widehat{\mathrm{Var}}_h = \sum_{j=1}^c (X_{(h,j)} - \bar{X}_h) \otimes (X_{(h,j)} - \bar{X}_h)/(c-1)$, where we use a double script (h, j) to denote the jth observation in the hth slice. Thus $\Gamma^{-1}E\left\{\Gamma - 2\mathrm{Var}(X|Y) + \mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y)\right\}$ can be estimated by $\widehat{\Gamma}_D^{-1}\left\{\widehat{\Gamma} - \frac{2}{H}\sum_{h=1}^H \widehat{\mathrm{Var}}_h + \frac{1}{H}\sum_{h=1}^H \widehat{\mathrm{Var}}_h \widehat{\Gamma}_D^{-1} \widehat{\mathrm{Var}}_h\right\}$.

Theorem 3. *Suppose*

$$\left\| E[\mathrm{Var}(X|Y)] - (1/H)\sum_{h=1}^H \widehat{\mathrm{Var}}_h \right\| = O_p(n^{-\gamma})$$

and

$$\left\| E[\mathrm{Var}(X|Y)\widehat{\Gamma}_D^{-1}\mathrm{Var}(X|Y)] - (1/H)\sum_{h=1}^H \widehat{\mathrm{Var}}_h \widehat{\Gamma}_D^{-1} \widehat{\mathrm{Var}}_h \right\| = O_p(t_D^{-1}n^{-\gamma})$$

for some $0 < \gamma \le 1/2$. *If* $D \to \infty$, $s_D = o_p(t_D)$, $1/(\sqrt{n}t_D^2) \to 0$, *and* $1/(n^\gamma t_D^{5/2}) \to 0$, *then* $\|\widehat{\beta}_j - \beta_j\| = o_p(1), j = 1, \ldots, K$.

Remark 4. Here we briefly discuss how polynomial rates for convergence to $E[\mathrm{Var}(X|Y)]$ and $E[\mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y)]$, as in the assumption of the theorem, can be obtained by the results in the existing literature. The rate of $\|E[\mathrm{Var}(X|Y)] - (1/H)\sum_{h=1}^H \widehat{\mathrm{Var}}_h\|$ for the multivariate case can be found in [23,36]. These results can be adapted easily to functional context using some necessary modifications that are contained in the proof of Proposition 2. In particular, under suitable assumptions, one can obtain $\|E[\mathrm{Var}(X|Y)] - (1/H)\sum_{h=1}^H \widehat{\mathrm{Var}}_h\| = O_p(n^{-1/2})$. Also, under assumptions similar to those in [26], and following their proof, we can show $\|E[\mathrm{Var}(X|Y)\widehat{\Gamma}_D^{-1}\mathrm{Var}(X|Y)] - \frac{1}{H}\sum_{h=1}^H \widehat{\mathrm{Var}}(h)\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}(h)\| = O_p(t_D^{-1}n^{-\gamma})$, for some $0 < \gamma \le 1/2$, where the main difference from the multivariate case is the appearance of t_D^{-1} which comes from $\|\widehat{\Gamma}_D^{-1}\|$ (since $s_D = o_p(t_D)$, $\|\widehat{\Gamma}_D^{-1}\|$ is also of order $O_p(t_D^{-1})$) that will come up in various places in the proof for the functional case. Note that in general we have $\gamma < 1/2$ for SAVE estimator. $\gamma = 1/2$ is possible if bias correction is adopted, or Y can take only a finite number of possible values, as shown in [26]. Our statement of the theorem directly impose these convergence rates as assumptions for clarity since list of those more primitive assumptions in [26] would be quite messy. Finally, the value of γ in $\|E[\mathrm{Var}(X|Y)] - (1/H)\sum_{h=1}^H \widehat{\mathrm{Var}}_h\|$ and $\|E[\mathrm{Var}(X|Y)\widehat{\Gamma}_D^{-1}\mathrm{Var}(X|Y)] - (1/H)\sum_{h=1}^H \widehat{\mathrm{Var}}_h\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}_h\|$ generally can be different. We assume them to be the same only for ease of notation. The convergence rate of $\|E[\mathrm{Var}(X|Y)\widehat{\Gamma}_D^{-1}\mathrm{Var}(X|Y)] - (1/H)\sum_{h=1}^H \widehat{\mathrm{Var}}_h\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}_h\|$ is usually slower and determines the convergence rate of $\widehat{\beta}_j$.

4. Implementation in finite dimension

The above presentation of the SIR and the SAVE estimator in terms of operator language does not make it clear how to implement the procedures in practice. In this section, we explain that we can project X_i onto S_D to obtain finite dimensional $x_i \in R^D$ and then perform SIR and SAVE similar to the multivariate case. Note that although in terms of implementation the functional case is similar to the multivariate case after a suitable basis is determined, in terms of theory the functional case is different from the multivariate case. In addition, when the basis is not orthonormal, some adjustments are necessary. Although conceptually the basis can be first transformed to be orthogonal by the Gram–Schmidt procedure, it is more convenient to directly work on the original basis, as usually done for polynomial splines.

Let $\mathcal{M} : L_2[0, 1] \to S_D$ be the mapping that maps any $\beta \in L_2[0, 1]$ to its best approximation (in the sense of minimizing L_2 norm) in the space S_D. By abuse of notation, $\mathcal{M}$ also maps any operator constructed from X to the operator constructed from $\mathcal{M}(X)$. For example, we have $\mathcal{M}(\Gamma) = E(x^TB \otimes x^TB)$, where $x^TB = \mathcal{M}(X)$ and $B = (B_1, \ldots, B_D)^T$.

To derive the computational algorithm, we first consider $\mathcal{M}(\Gamma)$. It is easy to see

$$\mathcal{M}(\Gamma)(b^TB) = E(\langle x^TB, b^TB\rangle x^TB) = (xx^T\bar{B}b)^TB = (\mathrm{Var}(x)\bar{B}b)^TB,$$

where $\bar{B}$ is the $D \times D$ matrix with entries $\langle B_d, B_{d'}\rangle = \int_0^1 B_d(t)B_{d'}(t)dt$. Thus based on the displayed equation above, with respect to the given basis, the operator $\mathcal{M}(\Gamma)$ is represented by the $D \times D$ matrix $\mathrm{Var}(x)\bar{B}$. Similarly, approximation to $\mathrm{Var}(E[X|Y])$ is represented by the matrix $\mathrm{Var}(E[x|Y])\bar{B}$. Thus the algorithm for computing a drs using SIR is to find the top K eigenvectors of $\bar{B}^{-1}\widehat{\mathrm{Var}(x)}^{-1}\widehat{\mathrm{Var}(E[x|Y])}\bar{B}$, denoted by $b_1, b_2, \ldots, b_K$ and then estimate β_k by b_k^TB, $k = 1, \ldots, K$, where $\widehat{\mathrm{Var}(x)}$ and $\widehat{\mathrm{Var}(E[x|Y])}$ are the estimators for $\mathrm{Var}(x)$ and $\mathrm{Var}(E[x|Y])$ respectively. Here we use $\widehat{\mathrm{Var}(x)} = \sum_{i=1}^n x_ix_i^T/n$ and $\widehat{\mathrm{Var}(E[x|Y])}$ is the usual slicing estimator used in finite dimensional sliced inverse regression. Note that after reducing to finite dimensions we could use other approaches, say kernel method [35], for $\widehat{\mathrm{Var}(E[x|Y])}$, but we only focus on the original proposal [24] in this paper.

Similarly, to use SAVE in functional context, we compute the top eigenvectors of $I_{D\times D} - 2\bar{B}^{-1}\widehat{\mathrm{Var}(x)}^{-1}\widehat{\mathrm{Var}(x|Y)}\bar{B} + \bar{B}^{-1}\widehat{\mathrm{Var}(x)}^{-1}\widehat{\mathrm{Var}(x|Y)}\widehat{\mathrm{Var}(x)}^{-1}\widehat{\mathrm{Var}(x|Y)}\bar{B}$, and use these as coefficients in the basis expansion to obtain the estimated drs. Although this expression looks quite complicated, when $\bar{B}$ is the identity matrix (as when we use functional PCA to estimate the basis) it is the same as the usual SAVE formula in finite dimensions.

5. Choice of dimension and regularization parameter

Determination of the dimension of the drs is a common problem in SIR and SAVE. In functional context, the regularization parameter D also needs to be determined. Determination of drs dimension can be done by looking at the sum of the minor eigenvalues of an appropriate matrix as suggested in [24,10]. In particular, we look at the eigenvalues of $\widehat{\mathrm{Var}}(E[X|Y])$ in SIR, and $\widehat{\Gamma} - \frac{2}{H}\sum_{h=1}^H \widehat{\mathrm{Var}}_h + \frac{1}{H}\sum_{h=1}^H \widehat{\mathrm{Var}}_h\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}_h$ in SAVE. If the asymptotic distribution for the test statistic can be derived, which is typically a mixture of chi-squared distributions as shown in the finite-dimensional case [24,10], test can be performed sequentially starting with the null hypothesis $K = 0$ (that is the null model where predictor does not have any effect on the response). However, for functional data, asymptotic distribution is harder to obtain, especially for SAVE. Thus we consider the permutation test, which is based on recalculating the test statistic using multiple random permutations of the responses. For detail, see Yin and Cook [33]. For SIR, we use the eigenvalues of $\widehat{\mathrm{Var}}(E[X|Y])$, so obviously no finite-dimensional approximation is necessary (in practice this means we can use any D for this testing step as long as it is large enough). For SAVE, we use the eigenvalues of $\widehat{\Gamma} - \frac{2}{H}\sum_{h=1}^H \widehat{\mathrm{Var}}_h + \frac{1}{H}\sum_{h=1}^H \widehat{\mathrm{Var}}_h\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}_h$ and we choose a relatively large $D = 7$ just to avoid singularity.

After the dimension is determined, there is still the problem of choosing D. Bias–variance tradeoff is controlled by this parameter. As D increases, variance increases and bias decreases.

For both choice of dimension of the drs and D, graphical methods that plot the responses versus the predictors using the estimated projections, or using some form of residuals, would be useful as discussed in [7,8]. Given that dimension reduction is often used for data exploration, the graphical method may be satisfactory enough.

Furthermore, if prediction is the ultimate goal, then the dimension of the drs as well as parameter D can be chosen by common methods used in regression, such as cross-validation. In fact, for choice of dimension, many multivariate regression methods are bundled with approaches for variable selection so dimension determination can be done in the context of multivariate regression, although these variable selection procedures typically do not respect the order of the predictors (in the context of dimension reduction using SIR or SAVE, it is desired that only the directions associated with larger eigenvalues are kept). One should note that many flexible nonparametric multivariate regression methods, including say neural networks or Gaussian process regression, suffer little from curse of dimensionality unless the dimension is very large. When these regression procedures are used for prediction, we can usually use a large number of directions obtained from SIR or SAVE.

6. Numerical examples

We use three simulation examples to illustrate functional dimension reduction. In all examples, the predictors are standard Brownian motion on [0, 1]. We set $n = 300$ and the predictor is observed without noise on a grid of equally spaced 100 points on [0, 1]. To study the sensitivity of results to the number of slices, we consider four values of H with $H = 5$, 10, 15 and 20. All the results presented are based on 100 simulated datasets in each scenario. We consider both functional PCA and B-spline basis. For the B-spline basis, the knots are chosen to be equally spaced on [0, 1]. For the latter we use the quadratic splines with various number of internal knots.

M1. $Y = \langle\beta_1, X\rangle + 100\langle\beta_2, X\rangle^3 + \epsilon$, $\beta_1(t) = \sin(3\pi t/2)$, $\beta_2(t) = \sin(5\pi t/2)$,
M2. $Y = \langle\beta_1, X\rangle^3 + 3\langle\beta_2, X\rangle + \epsilon$, $\beta_1(t) = (2t-1)^3 + 1$, $\beta_2(t) = (2t-1)^2 - 1$,
M3. $Y = 50\langle\beta_1, X\rangle\langle\beta_2, X\rangle^2 + \epsilon$, $\beta_1(t) = (2t-1)^2 - 1$, $\beta_2(t) = \sin(5\pi t/2)$,

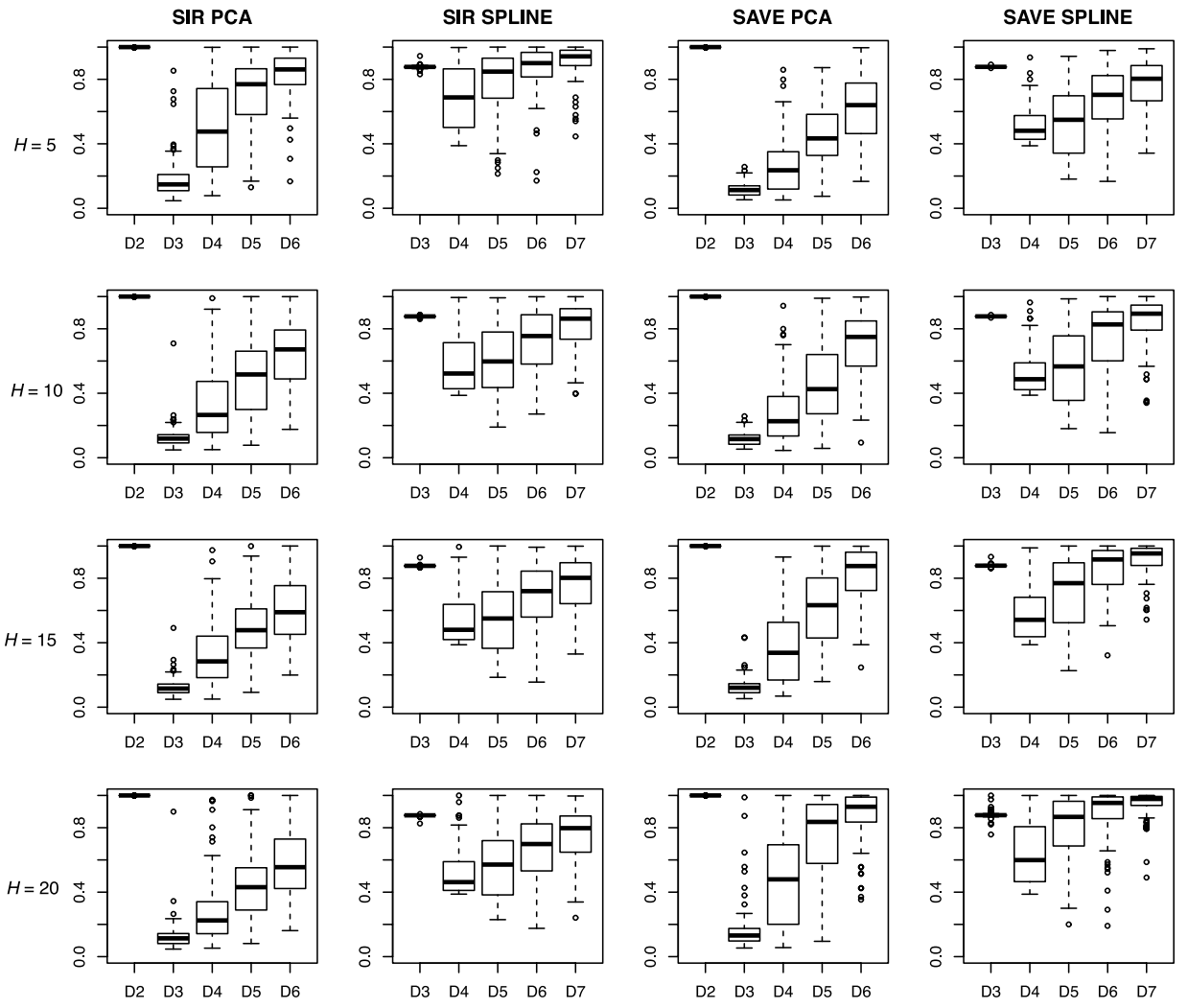

Fig. 1. Boxplots showing $\|P - \widehat{P}\|$ for M1 using four different estimation methods. Four rows correspond to four values of H, from small to large. Five boxplots in each case correspond to different values of D. For basis obtained from functional PCA, we present results for D from 2 to 6. For B-spline basis, we present results for the number of interval knots from 0 to 4 (D from 3 to 7).

where $\epsilon \sim N(0, 0.1^2)$. In M1, β_1 and β_2 are both the eigenfunctions of Brownian motion. M2 is similar to M1 but the directions are no long eigenfunctions of Brownian motion. M3 is intentionally designed such that SAVE will work better than SIR and is motivated by that SIR does not work well for quadratic or close to quadratic link functions.

We first consider the dimension $K = 2$ is known and study the accuracy of estimation. Let P and $\widehat{P}$ be the orthogonal projection operators onto the true drs and estimated drs respectively. The distance is measured by the largest singular value of $P - \widehat{P}$, denoted by $\|P - \widehat{P}\|$, with smaller values indicating better estimation performance. Figs. 1–3 show the boxplots of $\|P - \widehat{P}\|$ for M1–M3 using different values of D, for both SIR and SAVE. Here 5 values of D are used. Different rows in the figures correspond to different values of H. In general, we see that SIR and SAVE perform similarly, except for M3, for which SAVE is much better, as expected. For M1 and M3, functional PCA is better since some direction(s) are eigenfunctions of the covariance operator while for M2 the results for functional PCA and B-splines are similar. Finally, compared to the choice of D, the results are less sensitive to the choice of H. In the rest of the section, we only consider $H = 10$.

Next we present the permutation test results for M1–M3 (corresponding three different rows) in Figs. 4 and 5, where we use barplots to indicate the frequency of selection for different dimensions. We use a significance level of 0.05 for testing. In M1 and M2, permutation tests work well for SIR but tend to select $K = 1$ for SAVE. In M3, test for SIR only identifies one direction while test for SAVE identifies two directions most of the time. This is as expected since SIR does not work well for M3. Based on the simulation results, we suggest that permutation test can be used, but probably only as a rough guide. Graphical method or prediction-based choice as discussed previously could be used to supplement the tests findings.

Finally, we show that the graphical method can suggest reasonable value(s) for D. We only use model M1 with SIR method combined with basis estimated from functional PCA as an illustration. Fig. 6 shows the scatterplots of observations using two

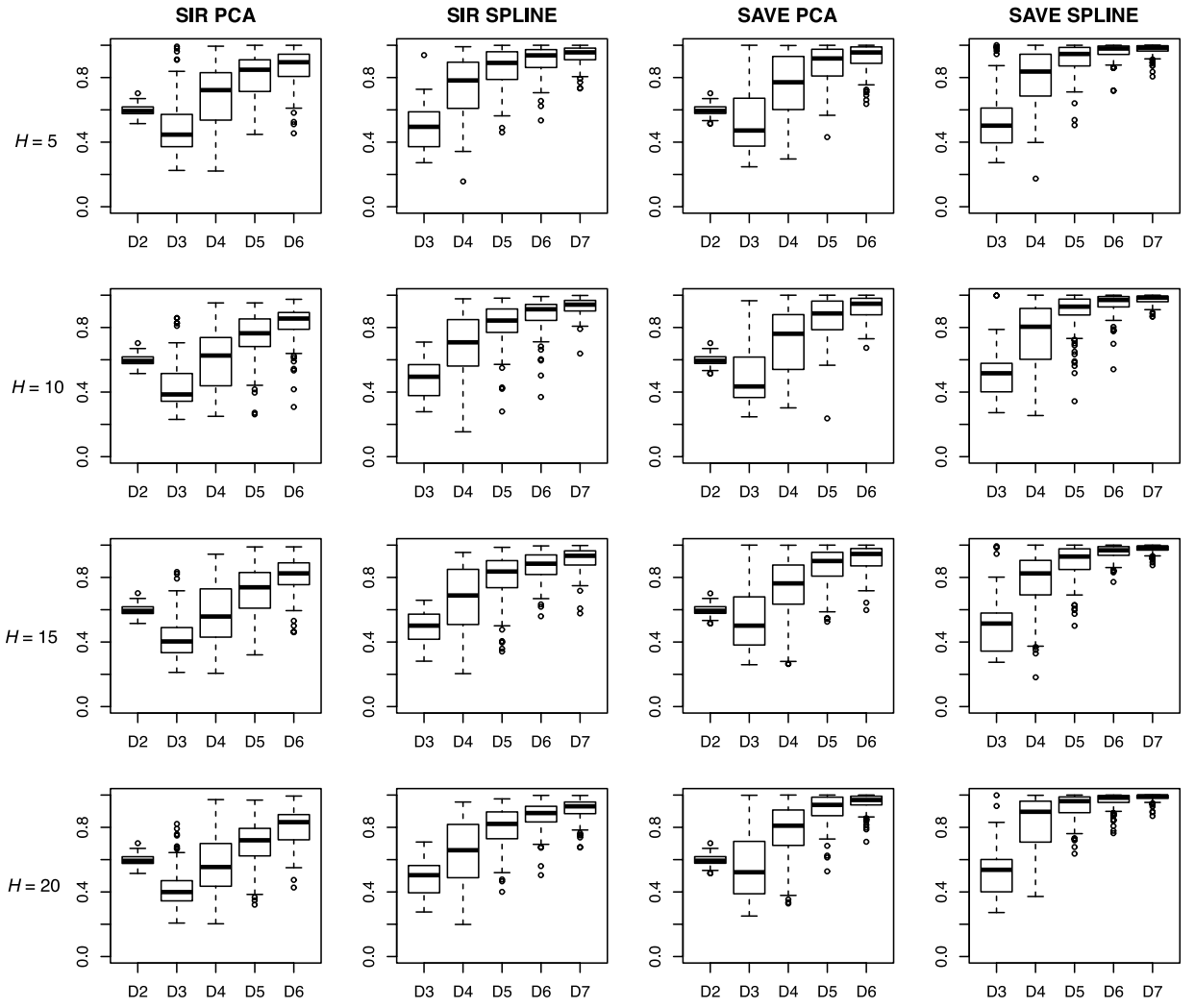

Fig. 2. Boxplots showing $\|P - \widehat{P}\|$ for M2.

estimated projections. Visually $D = 3$ (using three eigenvectors in functional PCA) seems to be slightly better in revealing the relationships between predictors and responses. More formally, regression can be performed to find the prediction errors, but we do not pursue the corresponding simulations here since we are mainly interested in recovering the directions per se. Prediction results will also depend on the regression method to be selected.

Finally, we use the Tecator spectrometric data which can be found at http://lib.stat.cmu.edu/datasets/tecator. For this dataset we used $H = 10$ since this is the value that turns out to results in the smallest errors after we tried $H = 5, 10, 15$. Each unit i (among $n = 215$) represents one piece of finely chopped meat. For each piece, we observe one spectrometric curve (X_i) which corresponds to the absorbance measured at 100 wavelengths. Moreover, for each piece we have its fat content (Y_i) obtained by an analytical chemical processing.

We applied both SIR and SAVE to this dataset. Permutation test suggests $K = 2$ or 3 is appropriate, and thus we only look at the first three eigenfunctions. We look at results for values of D from 3 to 7. For illustration purposes, here we consider prediction errors or different methods. We use the first 172 units for estimating the projection directions. For fitting the projection scores, we used Gaussian process regression (using *bgp* function in the R package *tgp* [21]) which is a flexible nonparametric regression method. The prediction errors on the test units are measured in root mean squared errors (RMSE). These errors are reported in Table 1. For this dataset, SIR generally performs better than SAVE and PCA better than splines. The smallest RMSE 0.90 is obtained by SIR using basis obtained by functional PCA.

For comparison in terms of prediction errors, we also compute the functional linear regression estimator by the functional principal component (PCR) approach [22] and the functional partial least squares (PLS) approach [14]. The number of basis considered are from 1 to 20. The prediction errors on the 45 test observations, measured by the absolute difference between the predicted response and the observed response, are show in Fig. 7. The smallest RMSE by PCR is 2.07, and the smallest

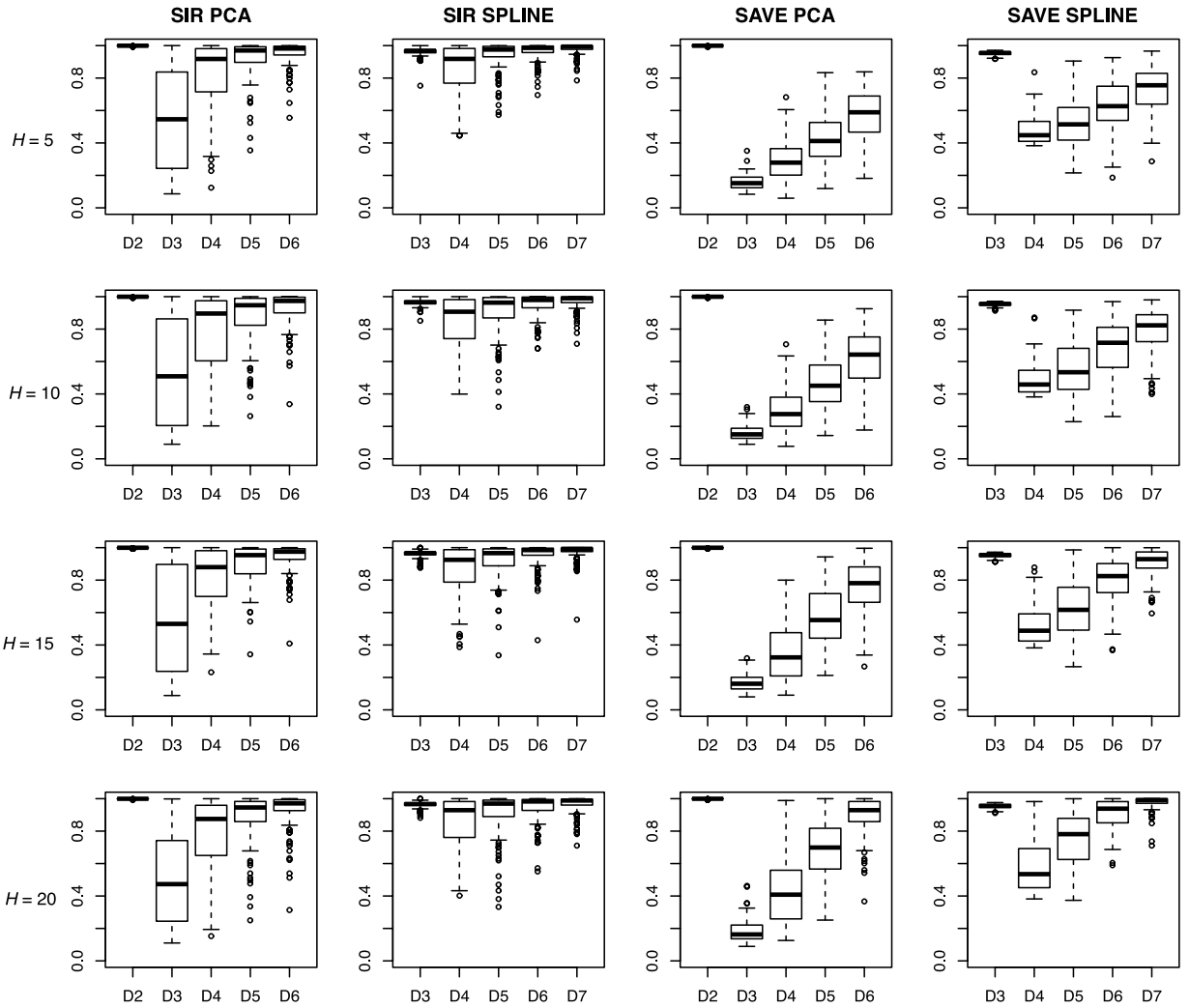

Fig. 3. Boxplots showing $\|P - \widehat{P}\|$ for M3.

Table 1
Prediction RMSE on the test data for the Tecator example.

	$D = 3$	$D = 4$	$D = 5$	$D = 6$	$D = 7$
SIR PCA	3.64	1.29	0.90	1.08	1.72
SIR SPLINE	3.58	1.39	1.42	2.71	4.00
SAVE PCA	4.29	1.68	1.70	1.13	2.10
SAVE SPLINE	3.76	2.11	1.71	2.99	3.94

RMSE by PLS is 1.77. This illustrates that functional linear regression using only one projection direction is not sufficient for this dataset.

7. Conclusion and discussion

In this paper, we studied SIR and SAVE for functional data and demonstrated their statistical consistency. Compared to existing works that focused on projecting the infinite-dimensional predictor onto subspaces derived from spectral decomposition of the covariance operator, as in functional PCA, we further extended the methodology to much more general basis systems. In the presentation we focused on basis obtained from spectral decomposition and B-spline basis only, but other basis such as wavelets can also be considered, which represents a very interesting class of decomposition method. Another important contribution is the study of functional SAVE which is not found in the existing literature. Obviously other dimension reduction methods can be extended to functional context, but we only focused on these two most common ones in this paper.

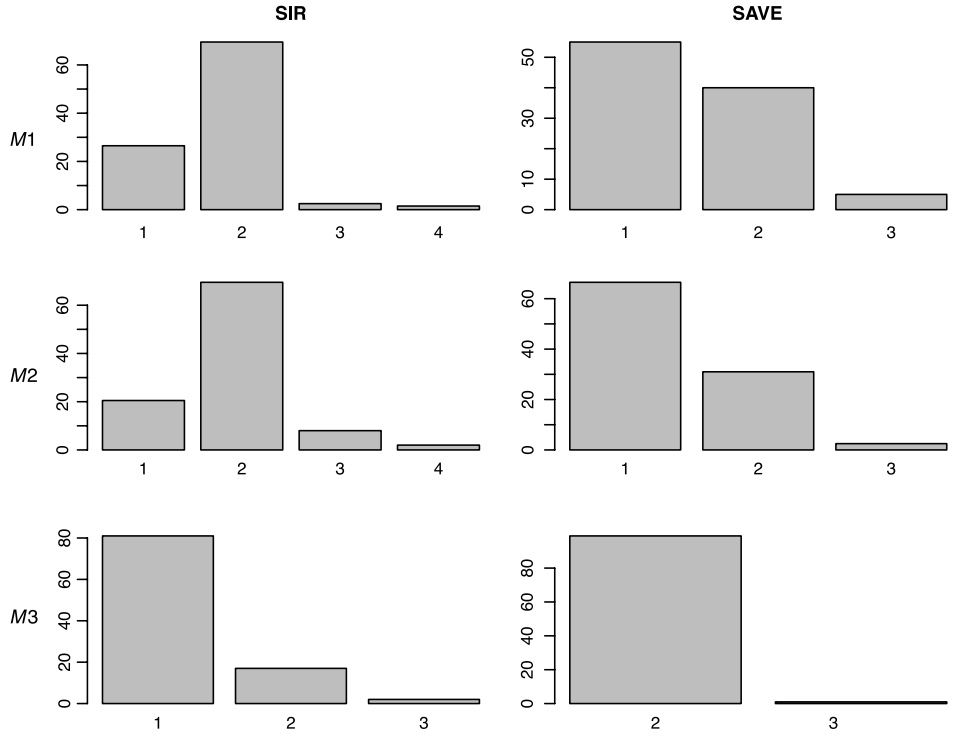

Fig. 4. Permutation test results for selection of the dimension of the drs, when using basis estimated by functional PCA.

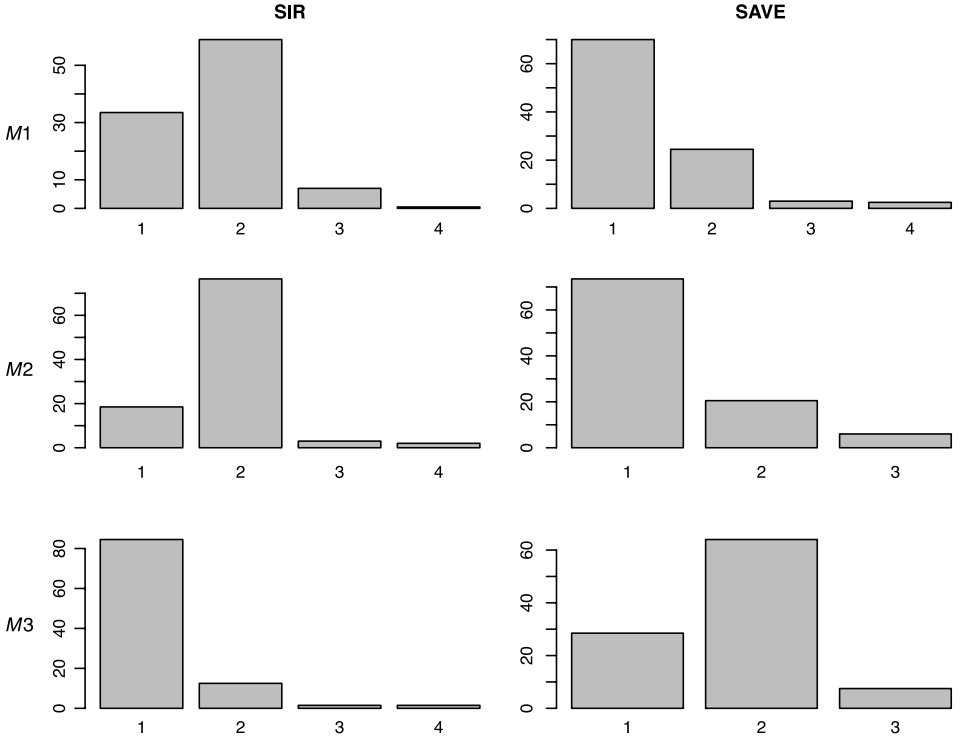

Fig. 5. Permutation test results for selection of the dimension of the drs, when using *B*-spline basis.

Some important and interesting issues are still left unresolved. The most noteworthy is the construction of better tests for determining the number of directions. This can be quite challenging for the SAVE estimator due to its complicated form. The

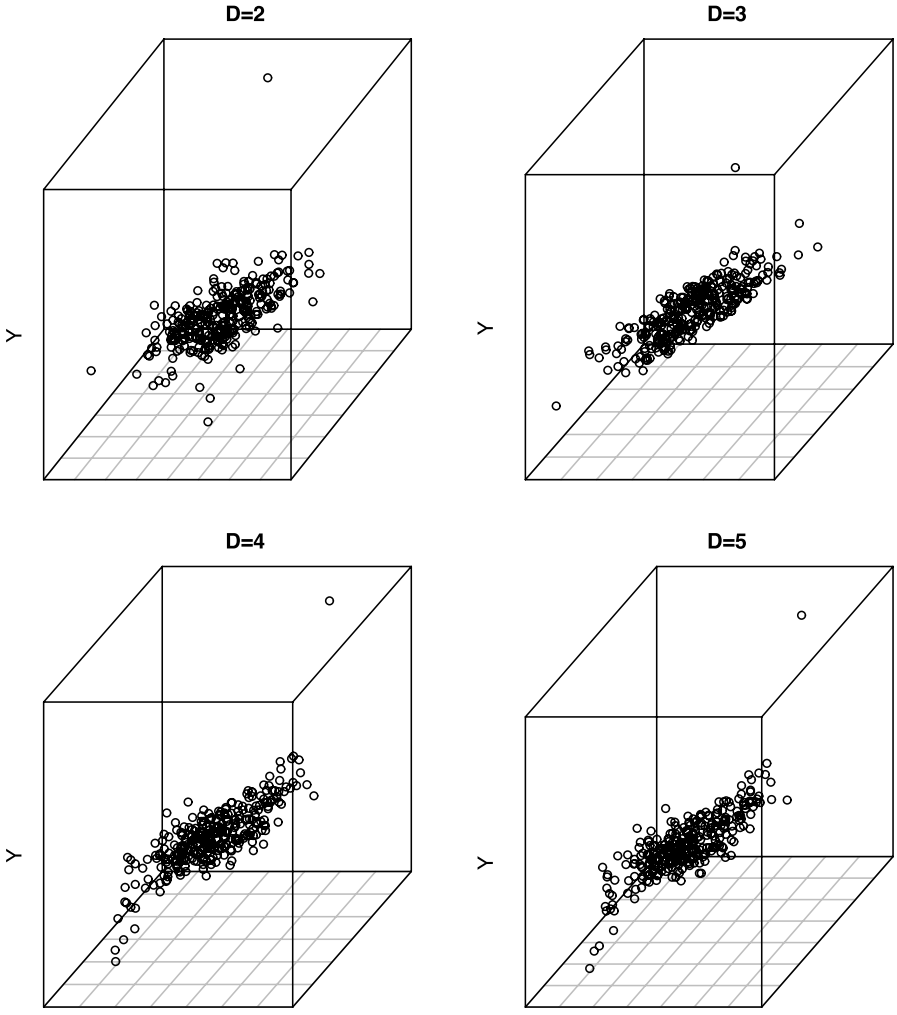

Fig. 6. 3D scatterplots using the two estimated projection directions, for four values of D.

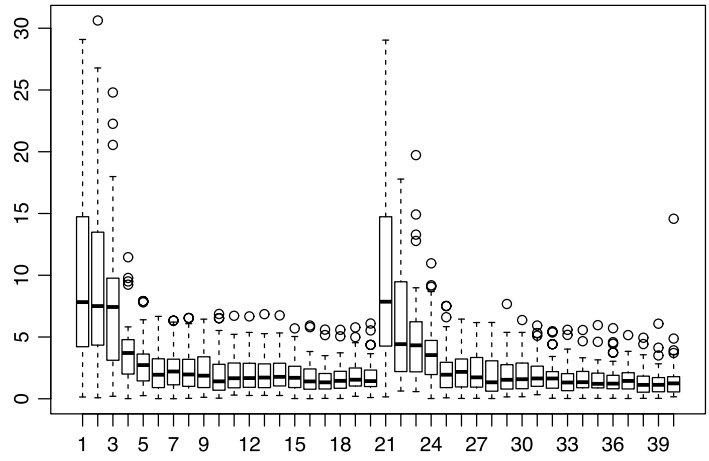

Fig. 7. Prediction errors for the Tecator data by functional principle component regression (20 boxes on the left) and functional partial least squares (20 boxes on the right).

graphical tools developed by Liquet and Saracco [27] could be adapted which can possibly select many parameters together, including H, K, D. A detailed numerical study is however outside the scope of the current paper. Another problem is to obtain

some nontrivial convergence rates. This would probably require stronger assumptions on the decay of the eigenvalues of the covariance operator, as in [22].

Acknowledgments

We sincerely thank two anonymous reviewers for their insightful comments that have led to significant improvement of the paper. The research of Heng Lian is supported by National Natural Science Foundation of China (11271241 and 11301279). Gaorong Li's research was supported by NSFC (11101014), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20101103120016), PHR (IHLB, PHR20110822), the Science and Technology Project of Beijing Municipal Education Commission (KM201410005010) and the Fundamental Research Foundation of Beijing University of Technology (X4006013201101).

Appendix. Proofs

In the proofs C denotes a generic positive constant.

Proof of Proposition 1. Note that for the first part, our assumption is less stringent than condition (A-3) in [19].

Using Karhunen–Loève expansion, we can write

$$\mathrm{Var}(E[X|Y]) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} E[E(\xi_i|Y)E(\xi_j|Y)]\phi_i \otimes \phi_j.$$

Thus for any $\beta = \sum_{i=1}^{\infty} b_i\phi_i \in L_2[0,1]$, we have

$$\mathrm{Var}(E[X|Y])\beta = \sum_{j=1}^{\infty}\left\{\sum_{i=1}^{\infty} b_i E[E(\xi_i|Y)E(\xi_j|Y)]\right\}\phi_j.$$

The domain of Γ^{-1} is $F_X = \{\beta = \sum_i b_i\phi_i \in L_2[0,1] : \sum_{i=1}^{\infty} b_i^2/\lambda_i^2 < \infty\}$. Thus we only need to show $\mathrm{Var}(E[X|Y])\beta \in F_X$ for all $\beta \in L_2[0,1]$, which is equivalent to showing $\sum_j\{\sum_i b_i E[E(\xi_i|Y)E(\xi_j|Y)]\}^2/\lambda_j^2 < \infty$. Since $\{\sum_i b_i E[E(\xi_i|Y)E(\xi_j|Y)]\}^2 \le \{\sum_i b_i^2\}\{\sum_i (E[E(\xi_i|Y)E(\xi_j|Y)])^2\}$ by Cauchy–Schwarz inequality, we see $\Gamma^{-1}\mathrm{Var}(E[X|Y])$ is well defined if $\sum_j \lambda_j^{-2}\{\sum_i E(E(\xi_i|Y)E(\xi_j|Y))^2\} < \infty$.

For the second part, we need to show that $\Gamma^{-1}E[\mathrm{Var}(X|Y)]$ and $\Gamma^{-1}E[\mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y)]$ are well-defined. Given any $\beta = \sum_i b_i\phi_i$, formally we have

$$\Gamma^{-1}E[\mathrm{Var}(X|Y)]\beta = \sum_j\left\{E\left[\sum_i \mathrm{Cov}_{ij|Y}b_i/\lambda_j\right]\right\}\phi_j,$$

which is well-defined if

$$\sum_j\left(E\left[\sum_i \mathrm{Cov}_{ij|Y}b_i/\lambda_j\right]\right)^2 < \infty.$$

Using the Cauchy–Schwarz inequality, we have

$$\sum_j\left(E\left[\sum_i \mathrm{Cov}_{ij|Y}b_i/\lambda_j\right]\right)^2 \le \sum_j E\left[\left(\sum_i \mathrm{Cov}_{ij|Y}b_i/\lambda_j\right)^2\right] \le \sum_j\left(E\sum_i \mathrm{Cov}_{ij|Y}^2\right)\left(\sum_i b_i^2\right)\Big/\lambda_j^2,$$

and thus $\Gamma^{-1}E[\mathrm{Var}(X|Y)]\beta$ is well-defined if

$$\sum_j\left(E\sum_i \mathrm{Cov}_{ij|Y}^2\right)\Big/\lambda_j^2 < \infty,$$

which is implied by (3).

Now denote $t_j = \sum_i \mathrm{Cov}_{ij|Y}b_i/\lambda_j$, and then we can write $\Gamma^{-1}\mathrm{Var}(X|Y)\beta = \sum_j t_j\phi_j$. Thus

$$\Gamma^{-1}\mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y)\beta = \sum_j\left(\sum_i \mathrm{Cov}_{ij|Y}t_i/\lambda_j\right)\phi_j,$$

and $\Gamma^{-1}E[\mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y)]$ is well-defined if

$$E\sum_j\left(\sum_i \mathrm{Cov}_{ij|Y}t_i/\lambda_j\right)^2<\infty. \tag{4}$$

Since

$$\sum_j\left(\sum_i \mathrm{Cov}_{ij|Y}t_i/\lambda_j\right)^2\le\sum_j\left(\sum_i \mathrm{Cov}^2_{ij|Y}\right)\left(\sum_i t_i^2\right)\Big/\lambda_j^2,$$

and

$$\sum_j t_j^2\le C\sum_j\left(\sum_i \mathrm{Cov}^2_{ij|Y}\right)\Big/\lambda_j^2,$$

(4) is also implied by (3). □

Proof of Proposition 2. Using $X_i=m(Y_i)+\epsilon_i$, we have

$$\begin{aligned}\widehat{\mathrm{Var}}(E[X|Y]) &= \frac{1}{H}\sum_{h=1}^{H}\left\{\frac{1}{c}\sum_{j=1}^{c}(m(Y_{(h,j)})+\epsilon_{(h,j)})\otimes\frac{1}{c}\sum_{j=1}^{c}(m(Y_{(h,j)})+\epsilon_{(h,j)})\right\}\\ &= \frac{1}{Hc^2}\sum_{h=1}^{H}\sum_{j=1}^{c}\sum_{l=1}^{c}m(Y_{(h,j)})\otimes m(Y_{(h,l)})+\frac{1}{Hc^2}\sum_{h=1}^{H}\sum_{j=1}^{c}\sum_{l=1}^{c}m(Y_{(h,j)})\otimes\epsilon_{(h,l)}\\ &\quad+\frac{1}{Hc^2}\sum_{h=1}^{H}\sum_{j=1}^{c}\sum_{l=1}^{c}\epsilon_{(h,j)}\otimes m(Y_{(h,l)})+\frac{1}{Hc^2}\sum_{h=1}^{H}\sum_{j=1}^{c}\sum_{l=1}^{c}\epsilon_{(h,j)}\otimes\epsilon_{(h,l)}\\ &=: D_{1n}+D_{2n}+D_{3n}+D_{4n}.\end{aligned}$$

We will show $\|D_{1n}-\mathrm{Var}(E[X|Y])\|=O_p(n^{-\gamma})$ and the other three terms are $O_p(n^{-\gamma})$. The proof strategy is similar to the proof for the finite-dimensional case, say Hsing and Carroll [23], although the latter focused on a different version of SIR. First, we have

$$\begin{aligned}D_{1n} &= \frac{1}{Hc^2}\sum_{h=1}^{H}\sum_{j=1}^{c}\sum_{l=1}^{c}m(Y_{(h,j)})\otimes m(Y_{(h,l)})\\ &= \frac{1}{Hc^2}\sum_{h=1}^{H}\sum_{j=1}^{c}\sum_{l=1}^{c}m(Y_{(h,j)})\otimes(m(Y_{(h,j)})+m(Y_{(h,l)})-m(Y_{(h,j)}))\\ &= \frac{1}{Hc}\sum_{i=1}^{n}m(Y_i)\otimes m(Y_i)+\frac{1}{Hc^2}\sum_{h,j\ne l}m(Y_{(h,j)})\otimes(m(Y_{(h,l)})-m(Y_{(h,j)}))\\ &=: D_{11n}+D_{12n}.\end{aligned}$$

Obviously $\|D_{11n}-\mathrm{Var}(E[X|Y])\|=O_p(n^{-1/2})$ by law of large numbers (note $\mathrm{Var}(E[X|Y])=E(m(Y)\otimes m(Y))$). To deal with D_{12n}, we divide the sum over h into three summations: from 1 to $[Hq]$, $[Hq]+1$ to $[H(1-q)]$, and $[H(1-q)]+1$ to H, with appropriately chosen small positive number q, and thus write $D_{12n}=D_{121n}+D_{122n}+D_{123n}$. For h from $[Hq]+1$ to $[H(1-q)]$, $Y_{(h,j)}$'s are contained in a bounded interval $[-T(q),T(q)]$ which implies

$$\begin{aligned}D_{122n} &= O_p\left(\frac{1}{Hc^2}\sum_{h=[Hq]}^{[H(1-q)]}\sum_{j<l}\|m(Y_{(h,j)})-m(Y_{(h,l)})\|\right)\\ &= O_p\left(\frac{1}{H}\sum_{i=c[Hq]+1}^{c[H(1-q)]}\|m(Y_{(i+1)})-m(Y_{(i)})\|\right)\\ &= o_p(n^{r-1}c),\end{aligned}$$

since m is bounded on compact sets.

For h from 1 to $[Hq]$ (the case for h from $[H(1-q)]+1$ to H is similar), if Y is unbounded, then we can find small enough q such that $Y_{(h,j)}\in(-\infty,T_0]$.

$$D_{121n}\le\frac{1}{Hc^2}\sum_{h=1}^{[Hq]}\sum_{j\ne l}\|m(Y_{(h,j)})\|\,\|m(Y_{(h,j)})-m(Y_{(h,l)})\|$$

$$\leq \frac{1}{Hc^2} \max_{1\leq i\leq n} \|m(Y_i)\| \sum_{h=1}^{[Hq]} \sum_{j\neq l} |M(Y_{(h,j)}) - M(Y_{(h,l)})|$$
$$= O_p\left(\frac{1}{H} \max_{1\leq i\leq n} \|m(Y_i)\| \, |M(Y_{(c[Hq])}) - M(Y_{(1)})|\right).$$

By Lemma A.1 in [23], $|M(Y_{(c[Hq])}) - M(Y_{(1)})| = o_p(n^{1/4})$. Using $E\|m(Y)\|^4 < \infty$, we have $P(\max_{1\leq i\leq n} \|m(Y_i)\| > an^{1/4}) \leq nP(\|m(Y_i)\| > an^{1/4}) \leq nE\|m(Y)\|^4/(an^{1/4})^4$ and thus $\max_{1\leq i\leq n} \|m(Y_i)\| = O_p(n^{1/4})$. Thus $D_{121n} = o_p(n^{1/4+1/4-1}c)$.

Next, for the multivariate case, $\|D_{2n}\| = O_p(n^{-1/2})$ can be shown by straightforward moment calculations conditional on order statistics $Y_{(1)}, \ldots, Y_{(n)}$. However, for functional predictor, we need to take some detour. We bound the operator norm $\|D_{2n}\|$ by the Hilbert–Schmidt norm, which is given by

$$\|D_{2n}\|_{HS}^2 = \sum_{m=1}^{\infty} \sum_{k=1}^{\infty} \langle D_{2n} e_m, e_k\rangle^2$$
$$= \sum_{m,k} \frac{1}{H^2c^4} \left[\sum_{h,j,l} \langle m(Y_{(h,j)}), e_m\rangle \langle \epsilon_{(h,l)}, e_k\rangle\right]^2,$$

where $\{e_j\}$ is any orthonormal basis of $L_2[0, 1]$. Then the conditional second moment of $\|D_{2n}\|_{HS}^2$ is

$$E[\|D_{2n}\|_{HS}^2 \, |Y_{(1)}, \ldots, Y_{(n)}] \leq \frac{1}{H^2c^4} \sum_{m,k,h} E\left[\left(\sum_{j,l} \langle m(Y_{(h,j)}), e_m\rangle \langle \epsilon_{(h,l)}, e_k\rangle\right)^2 |Y_{(1)}, \ldots, Y_{(n)}\right]$$
$$\leq \frac{1}{H^2c^4} \sum_{m,k,h} \left(\sum_{j} \langle m(Y_{(h,j)}), e_m\rangle\right)^2 \sum_{l} E\left[(\langle \epsilon_{(h,l)}, e_k\rangle)^2 | Y_{(1)}, \ldots, Y_{(n)}\right]$$
$$\leq \frac{C}{H^2c^2} \sum_{m,h,j} \langle m(Y_{(h,j)}), e_m\rangle^2$$
$$= \frac{C}{H^2c^2} \sum_{h,j} \|m(Y_{(h,j)})\|^2$$
$$= O_p(1/n),$$

by law of large numbers. Thus $\|D_{2n}\| = O_p(n^{-1/2})$. Similarly we can get $\|D_{3n}\| = O_p(n^{-1/2})$.

Finally, we have

$$D_{4n} = \frac{1}{Hc^2} \sum_{h=1}^{H} \sum_{j=1}^{c} \epsilon_{(h,j)} \otimes \epsilon_{(h,j)} + \frac{1}{Hc^2} \sum_{h,j\neq l} \epsilon_{(h,j)} \otimes \epsilon_{(h,l)}.$$

The first term on the right hand side above is $O_p(1/c)$ by law of large numbers, and the second term is $O_p(n^{-1/2})$ using similar argument as in the analysis of D_{2n}. □

Proof of Theorem 2. Let $\alpha_1, \ldots, \alpha_K$ be the K positive eigenvalues of $\Gamma^{-1}\mathrm{Var}(E[X|Y])$ and $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_K$ be the K eigenvalues of $\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}(E[X|Y])$. Later in the proof we will show that $B_n = o_p(1)$ (B_n will be defined later) which implies $|\alpha_j - \widehat{\alpha}_j| = o_p(1)$ and thus we can assume $\widehat{\alpha}_j, j = 1, \ldots, K$ are all positive and distinct. We only show the convergence of the first eigenvector $\widehat{\beta}_1$ since the proofs for all $\widehat{\beta}_j$'s are the same.

Before we continue, we note that in the following we will work with operators such as $\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}$. Strictly speaking, due to the appearance of $\Gamma^{-1/2}$, this operator is defined on $F_{XX} = \{b = \sum_i b_j\phi_i : \sum_i b_i^2/\lambda_i < \infty\}$ which is a dense subset of $L_2[0, 1]$. However, since $\mathrm{Var}(E[X|Y])$ is of finite-rank, it can be verified that $\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}$ is a closable operator and thus the domain can actually be extended to $L_2[0, 1]$. When we take $\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}$ to be the same as its closed extension, the operator is actually bounded. Thus all calculations in the proof work as in the usual case where appropriate operators are defined on $L_2[0, 1]$.

Let $\eta = \Gamma^{1/2}\beta_1$ and $\widehat{\eta} = \widehat{\Gamma}_D^{1/2}\widehat{\beta}_1$. We have $\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}\eta = \alpha_1\beta_1$ and $\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}(E[X|Y])\widehat{\Gamma}_D^{-1/2}\widehat{\eta} = \widehat{\alpha}_1\widehat{\beta}_1$. Thus

$$\|\widehat{\beta}_1 - \beta_1\| = \|\alpha_1^{-1}\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}\eta - \widehat{\alpha}_1^{-1}\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}(E[X|Y])\widehat{\Gamma}_D^{-1/2}\widehat{\eta}\|$$
$$\leq \widehat{\alpha}_1^{-1}\|\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2} - \widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}(E[X|Y])\widehat{\Gamma}_D^{-1/2}\| + \alpha_1^{-1}\|\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}\| \, \|\widehat{\eta} - \eta\|$$
$$+ \left|\alpha_1^{-1} - \widehat{\alpha}_1^{-1}\right| \|\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}\|.$$

Standard perturbation theory for self-adjoint operator implies

$$\|\widehat{\eta}-\eta\| \le C\|\Gamma^{-1/2}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}-\widehat{\Gamma}_D^{-1/2}\widehat{\mathrm{Var}}(E[X|Y])\widehat{\Gamma}_D^{-1/2}\|,$$

and

$$|\alpha_1-\widehat{\alpha}_1| \le \|\Gamma^{-1/2}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}-\widehat{\Gamma}_D^{-1/2}\widehat{\mathrm{Var}}(E[X|Y])\widehat{\Gamma}_D^{-1/2}\|.$$

Thus we have $\|\widehat{\beta}_1-\beta_1\| = O_p(A_n+B_n)$, where

$$A_n = \|\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}-\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}(E[X|Y])\widehat{\Gamma}_D^{-1/2}\|,$$
$$B_n = \|\Gamma^{-1/2}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}-\widehat{\Gamma}_D^{-1/2}\widehat{\mathrm{Var}}(E[X|Y])\widehat{\Gamma}_D^{-1/2}\|.$$

We will show $A_n = o_p(1)$. That $B_n = o_p(1)$ is similar.

Thus we consider A_n.

$$\begin{aligned} A_n &\le \|\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}-\Gamma_D^{-1}\mathrm{Var}(E[X|Y])\Gamma_D^{-1/2}\| + \|\Gamma_D^{-1}\mathrm{Var}(E[X|Y])\Gamma_D^{-1/2}-\widehat{\Gamma}_D^{-1}\mathrm{Var}(E[X|Y])\widehat{\Gamma}_D^{-1/2}\| \\ &\quad + \|\widehat{\Gamma}_D^{-1}(\mathrm{Var}(E[X|Y])-\widehat{\mathrm{Var}}(E[X|Y]))\widehat{\Gamma}_D^{-1/2}\| \\ &=: A_{1n}+A_{2n}+A_{3n}. \end{aligned}$$

For A_{1n}, we have

$$\begin{aligned} A_{1n} &\le \|(\Gamma^{-1}-\Gamma_D^{-1})\mathrm{Var}(E[X|Y])\Gamma^{-1/2}\| + \|\Gamma_D^{-1}\mathrm{Var}(E[X|Y])(\Gamma^{-1/2}-\Gamma_D^{-1/2})\| \\ &=: A_{11n}+A_{12n}. \end{aligned}$$

Since the range of $\mathrm{Var}(E[X|Y])\Gamma^{-1/2}$ is spanned by a finite number of elements $\{\Gamma\beta_1,\ldots,\Gamma\beta_K\}$, and it can be directly verified that $(\Gamma^{-1}-\Gamma_D^{-1})\Gamma\beta\to 0$ for any fixed β, we have $A_{11n}=o_p(1)$. Similarly, using that $\|A\|=\|A^*\|$ for any operator A with adjoint operator A^*, we have $A_{12n}=o_p(1)$.

For A_{2n}, we have

$$\begin{aligned} A_{2n} &\le \|(\Gamma_D^{-1}-\widehat{\Gamma}_D^{-1})\mathrm{Var}(E[X|Y])\Gamma_D^{-1/2}\| + \|\widehat{\Gamma}_D^{-1}\mathrm{Var}(E[X|Y])(\Gamma_D^{-1/2}-\widehat{\Gamma}_D^{-1/2})\| \\ &\le \|(\Gamma_D^{-1}-\widehat{\Gamma}_D^{-1})\mathrm{Var}(E[X|Y])\Gamma_D^{-1/2}\| + \|\Gamma_D^{-1}\mathrm{Var}(E[X|Y])(\Gamma_D^{-1/2}-\widehat{\Gamma}_D^{-1/2})\| \\ &\quad + \|(\Gamma_D^{-1}-\widehat{\Gamma}_D^{-1})\mathrm{Var}(E[X|Y])(\Gamma_D^{-1/2}-\widehat{\Gamma}_D^{-1/2})\| \\ &= A_{21n}+A_{22n}+A_{23n}. \end{aligned}$$

For A_{21n}, using the simple equality

$$A^{-1}-B^{-1} = B^{-1}(B-A)A^{-1}, \tag{5}$$

we have $(\Gamma_D^{-1}-\widehat{\Gamma}_D^{-1})\mathrm{Var}(E[X|Y])\Gamma_D^{-1/2} = \widehat{\Gamma}_D^{-1}(\widehat{\Gamma}-\Gamma)\Gamma_D^{-1}\mathrm{Var}(E[X|Y])\Gamma_D^{-1/2}$. It is easy to see that $\|\Gamma_D^{-1}\mathrm{Var}(E[X|Y])\Gamma_D^{-1/2}\| \le \|\Gamma^{-1}\mathrm{Var}(E[X|Y])\Gamma^{-1/2}\| < \infty$. Furthermore, since $\|\widehat{\Gamma}_D-\Gamma_D\| = o_p(t_D)$, we have $\|\widehat{\Gamma}_D^{-1}\| = O_p(t_D^{-1})$ and thus $A_{21n} = O_p(t_D^{-1}n^{-1/2}) = o_p(1)$.

To see $A_{22n}=o_p(1)$, we need to use the identify

$$A^{-1/2}-B^{-1/2} = A^{-1/2}(B^{3/2}-A^{3/2})B^{-3/2}+(A-B)B^{-3/2}, \tag{6}$$

and that

$$\|A^{3/2}-B^{3/2}\| \le C\|A-B\| \quad \text{if } A \text{ and } B \text{ are bounded linear operators}, \tag{7}$$

with constant C depending on the norm of A and B. The Eq. (6) can be directly verified while (7) can be found in Lemma 8 of Fukumizu et al. [20]. We then have

$$A_{22n} \le C\|t_D^{-1}(\widehat{\Gamma}-\Gamma)\| = O_p(n^{-1/2}t_D^{-1}) = o_p(1).$$

Using the identities (5) and (6), we can obtain $\|A_{23n}\|^2 = O_p(t_D^{-1}n^{-1}) = o_p(1)$ with similar arguments used for A_{21n} and A_{22n}.

Now we come to A_{3n}. Since $\|\mathrm{Var}(E[X|Y])-\widehat{\mathrm{Var}}(E[X|Y])\| = O_p(n^{-\gamma})$, we have $A_{3n} = O_p(n^{-\gamma}t_D^{-3/2}) = o_p(1)$. □

Proof of Theorem 3. The proof is similar to that for functional SIR with small modifications at the end. We can get $\|\widehat{\beta}_1-\beta_1\| = O_p(A_n+B_n)$, where $A_n = \|\Gamma^{-1}G\Gamma^{-1/2}-\widehat{\Gamma}_D^{-1}\widehat{G}\widehat{\Gamma}_D^{-1/2}\|$ and $B_n = \|\Gamma^{-1/2}G\Gamma^{-1/2}-\widehat{\Gamma}_D^{-1}\widehat{G}\widehat{\Gamma}_D^{-1/2}\|$. Here we denote $G = E\left\{\Gamma-2\mathrm{Var}(X|Y)+\mathrm{Var}(X|Y)\Gamma^{-1}\mathrm{Var}(X|Y)\right\}$ and $\widehat{G} = \widehat{\Gamma}-\frac{2}{H}\sum_{h=1}^{H}\widehat{\mathrm{Var}}_h+\frac{1}{H}\sum_{h=1}^{H}\widehat{\mathrm{Var}}_h\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}_h$.

Following the same arguments as for SIR, we only need to show that $\|\widehat{\Gamma}_D^{-1}(G_D-\widehat{G})\widehat{\Gamma}_D^{-1/2}\|=o_p(1)$, where $G_D=E\{\Gamma-2\mathrm{Var}(X|Y)+\mathrm{Var}(X|Y)\Gamma_D^{-1}\mathrm{Var}(X|Y)\}$. We use the decomposition

$$\begin{aligned}\|\widehat{\Gamma}_D^{-1}(G_D-\widehat{G})\widehat{\Gamma}_D^{-1/2}\| \le{}& \left\|\widehat{\Gamma}_D^{-1}\left(\Gamma-\widehat{\Gamma}-2\left(\mathrm{Var}(X|Y)-\frac{2}{H}\sum_{h=1}^{H}\widehat{\mathrm{Var}}_h\right)\right)\widehat{\Gamma}_D^{-1/2}\right\|\\ &+\|\widehat{\Gamma}_D^{-1}\mathrm{Var}(X|Y)(\Gamma_D^{-1}-\widehat{\Gamma}_D^{-1})\mathrm{Var}(X|Y)\widehat{\Gamma}_D^{-1/2}\|\\ &+\left\|\widehat{\Gamma}_D^{-1}\left(\mathrm{Var}(X|Y)\widehat{\Gamma}_D^{-1}\mathrm{Var}(X|Y)-\frac{1}{H}\sum_{h=1}^{H}\widehat{\mathrm{Var}}_h\widehat{\Gamma}_D^{-1}\widehat{\mathrm{Var}}_h\right)\widehat{\Gamma}_D^{-1/2}\right\|\\ =:{}& C_{1n}+C_{2n}+C_{3n}.\end{aligned}$$

Obviously, $C_{1n}=O_p(t_D^{-3/2}n^{-\gamma})=o_p(1)$. For C_{2n}, we can show $C_{2n}=\|\Gamma_D^{-1}\mathrm{Var}(X|Y)(\Gamma_D^{-1}-\widehat{\Gamma}_D^{-1})\mathrm{Var}(X|Y)\Gamma_D^{-1/2}\|+o_p(1)=O_p(\|\Gamma_D^{-1}-\widehat{\Gamma}_D^{-1}\|)+o_p(1)=o_p(1)$. Finally, $C_{3n}=O_p(t_D^{-5/2}n^{-\gamma})=o_p(1)$ by assumption. □

References

[1] A. Aguilera, F. Ocana, M. Valderrama, Estimation of functional regression models for functional responses by wavelet approximation, in: S. DaboNiang, F. Ferraty (Eds.), Functional and Operatorial Statistics, in: Contributions to Statistics, 2008, pp. 15–21.
[2] A. Ait-Saidi, F. Ferraty, R. Kassa, P. Vieu, Cross-validated estimations in the single-functional index model, Statistics 42 (6) (2008) 475–494.
[3] G. Aneiros-Perez, P. Vieu, Semi-functional partial linear regression, Statistics & Probability Letters 76 (11) (2006) 1102–1110.
[4] T.T. Cai, P. Hall, Prediction in functional linear regression, Annals of Statistics 34 (5) (2006) 2159–2179.
[5] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statistica Sinica 13 (3) (2003) 571–591.
[6] D. Chen, P. Hall, H. Müller, Single and multiple index functional regression models with nonparametric link, Annals of Statistics 39 (3) (2011) 1720–1747.
[7] R. Cook, On the interpretation of regression plots, Journal of the American Statistical Association 89 (425) (1994) 177–189.
[8] R. Cook, Graphics for regressions with a binary response, Journal of the American Statistical Association 91 (435) (1996) 983–992.
[9] R. Cook, L. Forzani, A. Yao, Necessary and sufficient conditions for consistency of a method for smoothed functional inverse regression, Statistica Sinica 20 (2010) 235–238.
[10] R. Cook, H. Lee, Dimension reduction in binary response regression, Journal of the American Statistical Association 94 (448) (1999) 1187–1200.
[11] R. Cook, L. Ni, Sufficient dimension reduction via inverse regression, Journal of the American Statistical Association 100 (470) (2005) 410–428.
[12] R. Cook, S. Weisberg, Sliced inverse regression for dimension reduction: comment, Journal of the American Statistical Association 86 (414) (1991) 328–332.
[13] C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators for functional linear regression, Annals of Statistics 37 (1) (2009) 35–72.
[14] A. Delaigle, P. Hall, Methodology and theory for partial least squares applied to functional data, The Annals of Statistics 40 (1) (2012) 322–352.
[15] N. Duan, K. Li, Slicing regression: a link-free regression method, The Annals of Statistics 19 (2) (1991) 505–530.
[16] F. Ferraty, P. Vieu, The functional nonparametric model and application to spectrometric data, Computational Statistics 17 (4) (2002) 545–564.
[17] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, in: Springer Series in Statistics, Springer, New York, NY, 2006.
[18] L. Ferré, A. Yao, Functional sliced inverse regression analysis, Statistics 37 (6) (2003) 475–488.
[19] L. Ferré, A. Yao, Smoothed functional inverse regression, Statistica Sinica 15 (3) (2005) 665–683.
[20] K. Fukumizu, F. Bach, A. Gretton, Statistical consistency of kernel canonical correlation analysis, The Journal of Machine Learning Research 8 (2007) 361–383.
[21] R. Gramacy, H. Lee, Bayesian treed Gaussian process models with an application to computer modeling, Journal of the American Statistical Association 103 (483) (2008) 1119–1130.
[22] P. Hall, J.L. Horowitz, Methodology and convergence rates for functional linear regression, Annals of Statistics 35 (1) (2007) 70–91.
[23] T. Hsing, R. Carroll, An asymptotic theory for sliced inverse regression, The Annals of Statistics 20 (2) (1992) 1040–1061.
[24] K. Li, Sliced inverse regression for dimension reduction, Journal of the American Statistical Association 86 (414) (1991) 316–327.
[25] B. Li, S. Wang, On directional regression for dimension reduction, Journal of the American Statistical Association 102 (479) (2007) 997–1008.
[26] Y. Li, L. Zhu, Asymptotics for sliced average variance estimation, The Annals of Statistics 35 (1) (2007) 41–69.
[27] B. Liquet, J. Saracco, A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches, Computational Statistics 27 (1) (2012) 103–125.
[28] C. Preda, Regression models for functional data by reproducing Kernel Hilbert spaces methods, Journal of Statistical Planning and Inference 137 (3) (2007) 829–840.
[29] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, second ed., in: Springer Series in Statistics, Springer, New York, 2005.
[30] H. Wong, R.Q. Zhang, W.C. Ip, G.Y. Li, Functional-coefficient partially linear regression model, Journal of Multivariate Analysis 99 (2) (2008) 278–305.
[31] S. Yang, General distribution theory of the concomitants of order statistics, The Annals of Statistics 5 (5) (1977) 996–1002.
[32] F. Yao, H.G. Müller, J.L. Wang, Functional linear regression analysis for longitudinal data, Annals of Statistics 33 (6) (2005) 2873–2903.
[33] X. Yin, R. Cook, Dimension reduction for the conditional *k*th moment in regression, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64 (2) (2002) 159–175.
[34] X. Yin, B. Li, R. Cook, Successive direction extraction for estimating the central subspace in a multiple-index regression, Journal of Multivariate Analysis 99 (8) (2008) 1733–1757.
[35] L. Zhu, K. Fang, Asymptotics for kernel estimate of sliced inverse regression, The Annals of Statistics 24 (3) (1996) 1053–1068.
[36] L. Zhu, K. Ng, Asymptotics of sliced inverse regression, Statistica Sinica 5 (1995) 727–736.
[37] L. Zhu, T. Wang, L. Ferré, Sufficient dimension reduction through discretization–expectation estimation, Biometrika 97 (2) (2010) 295–304.
[38] L. Zhu, L. Zhu, Z. Feng, Dimension reduction in regressions through cumulative slicing estimation, Journal of the American Statistical Association 105 (492) (2010) 1455–1466.