Statistica Sinica Preprint No: SS-2017-0277	
Title	Functional Sliced Inverse Regression in a Reproducing Kernel Hilbert Space: a Theoretical Connection to Functional Linear Regression
Manuscript ID	SS-2017-0277
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0277
Complete List of Authors	Guochang Wang and Heng Lian
Corresponding Author	Guochang Wang
E-mail	guochangwang66@hotmail.com
Notice: Accepted version subject to English editing.	

Functional Sliced Inverse Regression in a Reproducing Kernel Hilbert Space: a Theoretical Connection to Functional Linear Regression

Guochang Wang and Heng Lian

Jinan University and City University of Hong Kong

Abstract: We consider functional sliced inverse regression (FSIR) when the functional indices are assumed to be elements of a reproducing kernel Hilbert space. This work is motivated by the corresponding study in functional linear regression (FLR) in Cai and Yuan (2012), where a penalty involving the RKHS norm is used. Utilizing a close connection between FLR and FSIR that was not noted before, we show that FSIR can be dealt with by an analogy with FLR. Methodologically this is straightforward but the corresponding theoretical transfer from FLR to FSIR is nontrivial. In particular, we show that convergence rate for FSIR is the same as that for FLR and is thus minimax. This result is particularly interesting given the much more general specification of the dimension reduction problems compared to FLR. Simulations and real data are used to compare this with the functional PCA-based approach where the functional index is expanded by the eigenfunctions of the covariance kernel.

Key words and phrases: Convergence rate; Functional data; Sliced inverse regression.

1. Introduction: functional sliced inverse regression and functional linear regression

Dimension reduction for regression aims at reducing the dimension of a multivariate predictor X while preserving its predictive capability on a real-valued response Y (Li; 1991; Cook and Weisberg; 1991; Zhu and Fang; 1996; Cook and Lee; 1999; Yin and Cook; 2002; Cook and Ni; 2005). This class of approaches has been extended to the area of functional data analysis about which this article is concerned.

In the functional regression problem, let X be a square integrable random process indexed by $t \in [0,1]$ which is denoted simply by $X \in L_2[0,1]$, and let Y be a scalar random response. As always assumed in the functional linear regression (FLR) literature (Cardot et al.; 1999; Yao et al.; 2005; Cai and Hall; 2006; Hall and Horowitz; 2007), we assume $E\|X\|^4 < \infty$ where $\|X\| = (\int_0^1 X^2)^{1/2}$ is the L_2 norm of X. Without loss of generality, we also assume the predictor is centered with $EX = 0$. Functional dimension reduction seeks a set of square integrable functions, denoted by $\beta_1, \ldots, \beta_M$, such that Y depends on X only through the M inner products $\langle \beta_1, X\rangle, \ldots, \langle \beta_M, X\rangle$, where the inner product $\langle f, g\rangle = \int_0^1 fg$ for $f, g \in L_2[0,1]$. Mathematically, this can be formulated as $Y \perp X | (\langle \beta_1, X\rangle, \ldots, \langle \beta_M, X\rangle)$. That is, Y is independent of X given the M indices $\langle \beta_1, X\rangle, \ldots, \langle \beta_M, X\rangle$,

which means all information about Y in the process X is contained in the M-dimensional vector. Another way to formulate the problem is to pose it as a semiparametric regression problem

$$Y = g(\langle \beta_1, X \rangle, \ldots, \langle \beta_M, X \rangle, \epsilon),$$

where g is an unknown nonparametric link function and ϵ represents the noise in the regression problem. The M-dimensional subspace spanned by $\beta_1, \ldots, \beta_M$ (assuming they are linearly independent) is called the sufficient dimension reduction (sdr) space and denoted by $\mathcal{S}_{Y|X}$. The main objective is to estimate this space (instead of each specific direction which is unidentifiable in general). Note that the model above is very similar to the multiple index model. The difference is mainly that the former is more general while the latter imposes the more concrete additive error structure with $Y = g(\langle \beta_1, X \rangle, \ldots, \langle \beta_M, X \rangle) + \epsilon$. For example, in the model assumed for sufficient dimension reduction, the indices can affect both the mean and the variance. The multiple-index models are often estimated via more traditional approaches including kernels or series estimation, while sliced inverse regression only uses simple moment estimators. Sliced inverse regression (SIR), being the most commonly used dimension reduction estimator, has been extended to functional data (Ferré and Yao; 2003; Li and Hsing; 2010; Yao et al.; 2015).

The most popular method to obtain an estimator for either functional SIR (FSIR) or FLR is to use functional principal component analysis (F-PCA) which we now explain in this introduction. By Mercer's theorem, the covariance operator of the random process X, $\Gamma = E[X \otimes X]$, can be expressed as

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

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues and $\varphi_j \in L_2[0,1], j = 1, 2, \ldots,$ are the orthonormal set of eigenfunctions. Remember that for $f, g \in L_2[0,1]$, $f \otimes g$ is the linear operator that maps $h \in L_2[0,1]$ to $\langle g, h \rangle f \in L_2[0,1]$. Correspondingly, we have the Karhunen-Loève expansion $X = \sum_{j=1}^{\infty} \chi_j \varphi_j$ with $E\chi_j\chi_k = \lambda_j \delta_{jk}$ where $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ if $j \neq k$. We assume all the eigenvalues are strictly positive and distinct as usually imposed in the FLR and FSIR literature, which makes the estimation problem identifiable. Empirically, the eigenvalues and eigenfunctions can be estimated by the spectral decomposition of $\Gamma_n := \sum_{i=1}^{n} X_i \otimes X_i / n$ for i.i.d. data. To make the arguments slightly simpler, throughout the paper we also assume $\bar{X} := \sum_i X_i / n = 0$, otherwise we should define $\Gamma_n = \sum_{i=1}^{n} (X_i - \bar{X}) \otimes (X_i - \bar{X}) / n$, for example. The estimated eigenvalues and eigenfunctions will be denoted by $\{\hat{\lambda}_j, \hat{\varphi}_j\}$.

Illustrating the FPCA approach using FLR, one minimizes the ob-

jective functional $\sum_{i=1}^{n}(Y_i - \int \beta X_i)^2$ over all β that can be written as $\beta = \sum_{j=1}^{k} b_j \hat{\varphi}_j$ for some coefficients b_j. Note that the expansion is truncated at some finite integer k. It can be shown that the minimizer is $\hat{\beta} = (\hat{\Pi}_k \Gamma_n \hat{\Pi}_k)^+ (\sum_i X_i Y_i / n)$ where $\hat{\Pi}_k$ is the operator of projection onto the space spanned by $\hat{\varphi}_1, \ldots, \hat{\varphi}_k$, and $(.)^+$ denotes the pseudo-inverse. For FSIR, Ferré and Yao (2003) proved its consistency without making any connection to FLR, although their statement of consistency result hinted at a close similarity to FLR. Recovering this connection more explicitly is a nontrivial problem.

A crucial condition for the FPCA-based methods to work well is that the coefficient β in FLR (or indices in FSIR) can be efficiently represented in terms of the leading eigenfunctions of Γ, in the sense that the Fourier coefficients in the FPCA basis $\{\varphi_j\}$ decrease fast with j. As demonstrated in Cai and Yuan (2012) for FLR, this may not be true and there are opportunities for significant improvements. They proposed to solve the FLR problem by assuming the coefficient β lies in a known reproducing kernel Hilbert space (RKHS).

Motivated by Cai and Yuan (2012), one naturally wonders whether the methodological and theoretical results can be transferred to FSIR within the RKHS framework which can potentially improve on the FPCA-based

method for FSIR. Our theoretical approach to be laid out below is to transform the eigenvalue problem in $L_2[0,1]$ to a more standard eigenvalue problem in the Euclidean space, in the mean time studying the property of the new eigenvalue problem by uncovering a connection to FLR.

We believe our discovery of connections between FSIR and FLR is more generally applicable although here we only use estimation in the RKHS framework to illustrate that results in FLR can be transferred to FSIR. The rest of the article is organized as follows. In Section 2, we present the methodology of FSIR estimation in RKHS by making an informal connection to FLR, after a review of FSIR. We then present the asymptotic theory of our estimator for sdr space, which also relaxes some assumptions used in Cai and Yuan (2012). The proofs in the Appendix uncover a close relationship between FSIR and FLR which is key for proving the convergence rate of FSIR. In Section 4, simulations and a real dataset is used to show that the RKHS-based approach can provide improvements on the FPCA-based method for FSIR. Conclusion with discussions are made in Section 5. The technical proofs are relegated to the Appendix.

2. Methodology

2.1 FSIR based on FPCA

We herein review the basics of FSIR, mainly drawing results from Ferré and Yao (2003). Let $\Gamma\mathcal{S}_{Y|X}$ be the space spanned by $\Gamma\beta_1, \ldots, \Gamma\beta_M$. The principle of FSIR is based on the following result with proofs omitted, which is a direct extension of the multivariate case.

Proposition 1. *(Ferré and Yao; 2003) Suppose for all $b \in L_2[0,1]$, the conditional expectation $E[\langle b, X\rangle | \langle \beta_1, X\rangle, \ldots, \langle \beta_M, X\rangle]$ is linear in $\langle \beta_1, X\rangle, \ldots, \langle \beta_M, X\rangle$. Then $E(X|Y) \in \Gamma\mathcal{S}_{Y|X}$.*

The linearity condition in the proposition above constrains the marginal distribution of the predictors, not the conditional distribution of $Y|X$ as is typical in regression. It holds when X is a Gaussian process, although Gaussianity is not necessary.

The name of sliced inverse regression obviously originates from its use of $E[X|Y]$ instead of $E[Y|X]$. In functional linear regression it is assumed $E[Y|X] = \langle \beta, X\rangle$ for some $\beta \in L_2[0,1]$. Note that for simplicity we assumed there is no intercept in FLR, since the intercept can be estimated easily if necessary.

Based on Proposition 1, since $E[X|Y] \in \Gamma\mathcal{S}_{Y|X}$, we can estimate $\mathcal{S}_{Y|X}$

by estimating the eigenfunctions of $\Gamma^{-1}Var(E[X|Y])$, where $Var(E[X|Y]) = E[E(X|Y) \otimes E(X|Y)]$ is the covariance operator of $E[X|Y]$. Note if the eigenvalues of Γ are all positive, as typically assumed in the literature, Γ is invertible but Γ^{-1} is often not a bounded operator. Given i.i.d. data, as in FLR, Γ^{-1} can be estimated by $(\hat{\Pi}_k \Gamma_n \hat{\Pi}_k)^+$. To obtain the slicing estimator of $Var(E[X|Y])$, the range of Y is divided into H slices and we can estimate $Var(E[X|Y])$ by

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

where $\hat{X}_h$ is the sample average of the predictors whose associated response is in the hth slice.

From the descriptions above, it is suspected there is some connection between FLR and FSIR that makes it possible to transfer the asymptotic results proved on FLR to FSIR. In both cases, the functional PCA is used in calculating $(\hat{\Pi}_k \Gamma_n \hat{\Pi}_k)^+$. On the other hand, in FLR the coefficient β is obtained by applying $(\hat{\Pi}_k \Gamma_n \hat{\Pi}_k)^+$ to a random process $\sum_i X_i Y_i / n \in L_2[0,1]$, while in FSIR the object of interest is the eigenfunction of $(\hat{\Pi}_k \Gamma_n \hat{\Pi}_k)^+ \widehat{Var}(E[X|Y])$, making the connection unclear.

2.2 FSIR in a RKHS

Following Wahba (1990), a RKHS $\mathcal{H}$ is a Hilbert space of real-valued functions defined on, say, the interval $[0, 1]$, with inner product $\langle . \rangle_{\mathcal{H}}$, in which the point evaluation operator $L_t : \mathcal{H} \to R, L_t(f) = f(t)$ is continuous. The corresponding norm induced by the inner product is denoted by $\|.\|_{\mathcal{H}}$. By Riesz representation theorem, this definition implies the existence of a nonnegative definite, square integrable, bivariate function $K(s, t)$ such that $K(s, \cdot) \in \mathcal{H}$, and $\langle K(t, \cdot), f \rangle_{\mathcal{H}} = f(t)$ for every $f \in \mathcal{H}$ and $t \in [0, 1]$. To make the dependence on K explicit, the RKHS is denoted by $\mathcal{H}_K$ with the RKHS norm $\| \cdot \|_{\mathcal{H}_K}$. With abuse of notation, K also denotes the linear operator $f \in L_2 \to Kf = \int K(\cdot, s) f(s) ds$. For later use, we note that $\mathcal{H}_K$ is identical to the range of $K^{1/2}$.

For FLR, Cai and Yuan (2012) assumed that β is in a RKHS $\mathcal{H}_K$ and estimate β by

$$\hat{\beta} = \arg \min_{\beta \in \mathcal{H}_K} \sum_i (Y_i - \int X_i \beta)^2 + n\lambda \|\beta\|^2_{\mathcal{H}_K}, \tag{2.1}$$

where $\|.\|_{\mathcal{H}_K}$ is the RKHS norm and λ is a tuning parameter for the penalty. It was demonstrated that when the covariance kernel Γ does not align with the reproducing kernel K, the estimate obtained in RKHS can be much more accurate.

As mentioned in the introduction, the covariance operator is $\Gamma = EX \otimes X$ and we also use Γ to denote the covariance kernel $\Gamma(s,t) = EX(s)X(t)$. Perfect alignment between K and Γ means that the eigenfunctions ordered by the magnitudes of the eigenvalues are exactly the same for the two kernels/operators. Without assuming the two are aligned, Cai and Yuan (2012) used (2.1) to find the estimator of β. Noting that $\beta \in \mathcal{H}_K$ is equivalent to $\beta = K^{1/2}f$ for some $f \in L_2[0,1]$, and using the property $\|\beta\|_{\mathcal{H}_K} = \|f\|$, (2.1) is equivalent to

$$\arg \min_{f \in L_2[0,1]} \sum_i (Y_i - \langle K^{1/2}X_i, f\rangle)^2 + n\lambda \|f\|^2,$$

with solution $\hat{f} = (T_n + \lambda I)^{-1}(\sum_i K^{1/2}X_iY_i/n)$ where $T_n = K^{1/2}\Gamma_n K^{1/2}$ and I is the identity operator. For the population version, the solution to $\arg\min_f E(Y - \langle K^{1/2}X, f\rangle)^2$ is $T^{-1}E[K^{1/2}XY]$ where $T = K^{1/2}\Gamma K^{1/2}$. Informally, the above displayed equation means that we can simply replace X by $K^{1/2}X$ and focus on estimation of $f = K^{-1/2}\beta \in L_2[0,1]$ and the estimation of f does not involve consideration of RKHS any more.

Based on this observation, we can construct FSIR estimator in a RKHS by replacing X with $K^{1/2}X$. Assume that elements in $\mathcal{S}_{Y|X}$ are contained in $\mathcal{H}_K$. Let $\mathcal{S}^*_{Y|X} = K^{-1/2}\mathcal{S}_{Y|X} = \{f : f = K^{-1/2}\beta \text{ for some } \beta \in \mathcal{S}_{Y|X}\}$. Since $E[X|Y] \in \Gamma\mathcal{S}_{Y|X}$, we have $E[K^{1/2}X|Y] \in K^{1/2}\Gamma\mathcal{S}_{Y|X} = T\mathcal{S}^*_{Y|X}$ where $T = K^{1/2}\Gamma K^{1/2}$ and thus $\mathcal{S}^*_{Y|X}$ can be estimated by the space spanned

by the eigenfunctions of $T^{-1}Var(E[K^{1/2}X|Y])$. We summarize the above arguments in the following proposition.

Proposition 2. *Suppose $\beta_1, \ldots, \beta_K$ are in $\mathcal{H}_K$, and that for all $b \in L_2[0,1]$ the conditional expectation $E[\langle b, X\rangle | \langle \beta_1, X\rangle, \ldots, \langle \beta_M, X\rangle]$ is linear in $\langle \beta_1, X\rangle, \ldots, \langle \beta_M, X\rangle$. Then $E[K^{1/2}X|Y] \in T\mathcal{S}^*_{Y|X}$, and the eigenfunctions of $T^{-1}Var(E[K^{1/2}X|Y])$ associated with its nonzero eigenvalues are inside $\mathcal{S}^*_{Y|X}$.*

Empirically, given i.i.d. data, $T^{-1}Var(E[K^{1/2}X|Y])$ is estimated by $(T_n + \lambda I)^{-1}\widehat{Var}(E[K^{1/2}X|Y])$ where

$$\widehat{Var}(E[K^{1/2}X|Y]) = \frac{1}{H}\sum_{h=1}^{H}(K^{1/2}\hat{X}_h) \otimes (K^{1/2}\hat{X}_h).$$

For simplicity of asymptotic analysis in the next section, following the literature of sliced inverse regression, we assume Y is discrete taking only a finite number of values $y_1, \ldots, y_H$, with probabilities $p_1, \ldots, p_H$. This kind of simplification is typically used in the SIR literature, including Li (1991); Duan and Li (1991); Cook and Ni (2005). As argued in Cook and Ni (2005), 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 $\mathcal{S}_{\tilde{Y}|X} \subseteq \mathcal{S}_{Y|X}$, and when H is sufficiently large these two dimension reduction spaces are equal. Thus assuming Y is discrete does not lose much generality, and Ferré and Yao

(2003) also made this assumption. Thus we can write

$$Var(E[K^{1/2}X|Y]) = \sum_{h=1}^{H} p_h E[K^{1/2}X|Y = y_h] \otimes E[K^{1/2}X|Y = y_h],$$

which can thus be estimated by

$$\widehat{Var}(E[K^{1/2}X|Y]) = \sum_{h=1}^{H} \widehat{p}_h (K^{1/2}\hat{X}_h) \otimes (K^{1/2}\hat{X}_h),$$

where $\hat{X}_h$ is the average of X_i in the hth slice defined by $D_h = \{i : Y_i = y_h\}$ and $\widehat{p}_h = |D_h|/n$.

3. Convergence rate of FSIR estimator in a RKHS

Given the FSIR estimator constructed in the previous section simply by replacing X with $K^{1/2}X$, it is still unclear that FSIR can achieve the same rate of convergence as FLR in a RKHS. Let $\hat{f}_j, j = 1, \ldots, M$ (with $\|\hat{f}\| = 1$) be the eigenfunctions of $(T_n + \lambda I)^{-1}\widehat{Var}(E[K^{1/2}X|Y])$ associated with its top M eigenvalues, and let $\hat{\beta}_j = K^{1/2}\hat{f}_j$. The following technical assumptions are imposed.

(A1) $E\|X\|^4 < \infty$. Y is discrete taking H values $y_1, \ldots, y_H$. Both the reproducing kernel K and the covariance kernel Γ are positive definite.

(A2) Suppose the spectral expansion of T is $T = \sum_j s_j \psi_j \otimes \psi_j$. Note T is just the covariance operator when the predictor is $K^{1/2}X$. Recall

the Karhunen-Loéve expansion $K^{1/2}X = \sum_{j\geq 1}\xi_j\psi_j$. There exists a constant c such that $E[\xi_j^4] \leq c(E[\xi_j^2])^2$ for all $j \geq 1$.

(A3) There exists a positive, convex, decreasing function $\phi : (0,\infty) \to R^+$ with $\lim_{x\to\infty}\phi(x) = 0$, such that $s_j = \phi(j)$ at least for large j.

(A4) The operator $T^{-1}Var(E[K^{1/2}X|Y])$ has M eigenfunctions $f_1,\ldots,f_M$ (with $\|f_j\| = 1$) associated with the distinct eigenvalues $\alpha_1 > \cdots > \alpha_M > 0$. $\mathcal{S}^*_{Y|X}$ is spanned by $f_1,\ldots,f_M$ and thus $\mathcal{S}_{Y|X}$ is spanned by $K^{1/2}f_1,\ldots,K^{1/2}f_M$.

Assumption (A1) imposes a mild moment condition on the predictor typically assumed in the FLR and FSIR literature. Assumption of positive definiteness of Γ is necessary for identifiability (otherwise we can only estimate the component of β inside the space orthogonal to the kernel space of Γ. As in Cai and Yuan (2012), positive definiteness of K is mainly used for theoretical convenience. Assumption (A2) is similar to that assumed in Hall and Horowitz (2007); Cardot et al. (2007). Cai and Yuan (2012) assumed that $E(\int X(t)f(t)dt)^4 \leq c(E(\int X(t)f(t)dt)^2)^2$ for all $f \in L_2[0,1]$. This assumption implies (A2) which can be seen by choosing $f = K^{1/2}\psi_j$. Assumption (A3) also appeared in Cardot et al. (2007). Cai and Yuan (2012) considered a much more restrictive polynomial decay assumption

$s_j \asymp j^{-2r}$ for some $r > 0$, which corresponds to $\phi(x) = x^{-2r}$. Taking $\phi(x) = c_1 e^{-c_2 x}$ for some constants $c_1, c_2 > 0$, exponential decay of eigenvalues is also a special case of our result, among many others. Eigenvalues of K that decay at a rate j^{-2r} are more common. Among other examples, this type of scaling covers the case of Sobolev spaces, say consisting of functions with r derivatives (Birman and Solomjak; 1967; Raskutti et al.; 2012). A prominent kernel with exponentially decaying eigenvalues is the Gaussian kernel (Rasmussen and Williams; 2006). When $K = \Gamma$, it is clear that $T = K^{1/2}\Gamma K^{1/2}$ also has the polynomially or exponentially decaying eigenvalues. In more general cases with $K \neq \Gamma$, concrete examples seem much harder to construct. Referring to Proposition 2, (A4) merely assumes that in the population FSIR can recover the entire sdr space. This assumption is not necessary and is used for convenience of exposition. In general, the span of eigenfunctions extracted from $T^{-1}Var(E[K^{1/2}X|Y])$ is only a subspace of $\mathcal{S}^*_{Y|X}$. In this case, what one can show is only the convergence of the estimated $\hat{f}_j$ to the true eigenfunctions f_j, whose span is not the whole sdr space. In this case, of course there is no hope of recovering the whole sdr space in general using FSIR.

The risk measure we consider is the prediction risk $E^*(\langle\hat{\beta}_j, X^*\rangle - \langle\beta_j, X^*\rangle)^2$ where X^* is a copy of X independent of the training data and E^* is the

expectation taken over X^*. This risk is more natural than $\|\hat{\beta}_j - \beta_j\|$ since in FSIR typically we use $X_i\hat{\beta}_j$ either to plot them against Y_i for data exploration, or to treat them as the new predictors in multivariate regression. Since $\hat{\beta}_j = K^{1/2}\hat{f}_j$ and $\beta_j = K^{1/2}f_j$, this risk can also be written as $\|T^{1/2}(\hat{f}_j - f_j)\|^2$, where $T^{1/2}$ is the square root of T (that is $T^{1/2}T^{1/2} = T$).

Theorem 1. *Under assumptions (A1)-(A4), and taking λ to be the solution of $n\lambda = \phi^{-1}(\lambda)$, then for each $j \in \{1, \ldots, M\}$, there exists $c_j \in \{-1, 1\}$ such that*

$$E^*(c_j\langle\hat{\beta}_j, X^*\rangle - \langle\beta_j, X^*\rangle)^2 = O_p\left(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}\right),$$

uniformly for models with $\beta \in \mathcal{H}_K, \|\beta\|_{\mathcal{H}_K} = 1$. More specifically, the uniform upper bound by definition means that

$$\lim_{a\to\infty}\lim_{n\to\infty}\sup_{\|\beta\|_{\mathcal{H}_K}=1}\min_{c_j\in\{-1,1\}} P\left(E^*(c_j\langle\hat{\beta}_j, X^*\rangle - \langle\beta_j, X^*\rangle)^2 \geq a\left(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}\right)\right) = 0.$$

Since the eigenfunctions are only identifiable up to a sign change, c_j is necessary in the above to show the convergence rate.

Roughly speaking, in the convergence rate, λ represents the squared bias and $\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}$ represents the variance. The λ that satisfies $n\lambda = \phi^{-1}(\lambda)$ is actually chosen to trade off these two terms to make them of the same order. Thus the convergence rate is actually $O_p(\lambda)$ for this λ. We leave both two terms in the statement of the theorem to make the bias and variance

more explicit. To see that this λ balances the two terms in the rate above, let $J = \lfloor \phi^{-1}(\lambda) \rfloor$ be the integer part of $\phi^{-1}(\lambda)$. By splitting the sum over j into $j \leq J$ and $j > J$, we have

$$\frac{1}{n} \sum_j \frac{s_j^2}{(s_j + \lambda)^2} \leq \frac{J}{n} + \frac{s_{J+1} \sum_{j \geq J+1} s_j}{n\lambda^2}.$$

Since λ is the solution to the equation

$$\phi^{-1}(\lambda) = n\lambda, \tag{3.2}$$

we have $J = \lfloor \phi^{-1}(\lambda) \rfloor \leq \phi^{-1}(\lambda)$ and

$$\frac{s_{J+1} \sum_{j \geq J+1} s_j}{n\lambda^2} \leq \frac{(J+2) s_{J+1}^2}{n\lambda^2} \leq \frac{J+2}{n},$$

where we used that $\sum_{j \geq J+1} s_j \leq (J+2) s_{J+1}$ obtained from Lemma 1 of Cardot et al. (2007), and that $s_{J+1} = \phi(J+1) \leq \phi(\phi^{-1}(\lambda)) = \lambda$ by the definition of J. Thus we have

$$E^*(c_j \langle \hat{\beta}_j, X^* \rangle - \langle \beta_j, X^* \rangle)^2 = O_p(\lambda),$$

with λ defined by (3.2), which characterizes the optimal convergence rate. In the special case $\phi(x) = x^{-2r}$, $\lambda = n^{-2r/(2r+1)}$, which is the same as the rate obtained in Cai and Yuan (2012) for FLR. On the other hand, if $\phi(x) = e^{-x}$, we can easily show that $\log\log n/n < \lambda < \log n/n$, an almost parametric rate. Finally, for future reference, we note that by the property assumed for ϕ, it is easy to see that λ obtained from (3.2) satisfies $\lambda \to 0, \lambda n \to \infty$.

We now establish the lower bound. Obviously the lower bound for the special case that the true model is FLR with $Y = \langle \beta, X \rangle + \epsilon$ where X is a Gaussian process with a positive definite kernel Γ, $\|K^{-1/2}\beta\| = 1$, $\epsilon \sim N(0, \sigma^2)$ provides a lower bound for FSIR. Indeed, in this case, we can easily see that $E[X|Y] \neq 0$ (since (X, Y) are jointly Gaussian and nondegenerate). Thus $\mathcal{S}_{Y|X}$ is spanned by a single element β and $T^{-1}Var(E[K^{1/2}X|Y])$ has one nonzero eigenvalue with the corresponding eigenfunction exactly $K^{-1/2}\beta$. The lower bound for FLR has been considered in Cai and Yuan (2012) before. A slightly different construction is necessary here to deal with more general ϕ. The details of the proof are contained in the Appendix.

Theorem 2. *Consider the FLR with i.i.d. data: $Y_i = \langle \beta, X_i \rangle + \epsilon_i, i = 1, \ldots, n$. Given positive definite kernel K, and covariance operator Γ, suppose the eigenvalues $\{s_j\}$ of $T = K^{1/2}\Gamma K^{1/2}$ satisfy $s_j = \phi(j)$ for a positive, convex, decreasing function ϕ and let λ be defined by (3.2). Then, for any $a > 0$*

$$\lim_{a \to 0} \lim_{n \to \infty} \inf_{\hat{\beta}} \sup_{\beta \in \mathcal{H}_K, \|\beta\|_{\mathcal{H}_K} = 1} P(E^*(\langle \hat{\beta}, X^* \rangle - \langle \beta, X^* \rangle)^2 > a\lambda) = 1,$$

where the infimum is taken over all possible estimators based on the training data $(X_i, Y_i), i = 1, \ldots, n$. If the response Y_i is discretized to generate $\tilde{Y}_i$, the lower bound of course still holds for any estimator based on $(X_i, \tilde{Y}_i)$ since an estimator based on $(X_i, \tilde{Y}_i)$ is also an estimator based on (X_i, Y_i).

4. Numerical Results

4.1 Simulations

The purpose of this simulation is to compare the FPCA method of Ferré and Yao (2003) and the RKHS method for FSIR. A main message is that the methodological transfer from FLR to FSIR result in a very similar improvement to the FPCA-based approach. We use two simulation examples. The first simulation setup is similar to that used in Cai and Yuan (2012). We consider the RKHS with kernel

$$K(s,t) = \sum_{j\geq 1} \frac{2}{(j\pi)^4} \cos(j\pi s)\cos(j\pi t),$$

and thus $\mathcal{H}_K$ consists of functions of the form

$$f(t) = \sum_{j\geq 1} f_j \cos(j\pi t)$$

such that $\sum_j j^4 f_j^2 < \infty$. In this case, we actually have $\|f\|^2_{\mathcal{H}_K} = \int (f'')^2$.

We generate the data from the model

$$Y_i = \exp\{\langle \beta_1, X_i\rangle/5\} \cdot \langle \beta_2, X_i\rangle + \epsilon_i, i = 1, \ldots, n,$$

where $\beta_1(t) = \sum_{j=1}^{50} (4\sqrt{2}(-1)^j/j^2)\cos(j\pi t)$ and $\beta_2(t) = -2\sqrt{2}\cos(\pi t) - 4\sqrt{2}\cos(2\pi t) + 9\sqrt{2}\cos(3\pi t)$. The noises are generated from $N(0,\sigma^2)$.

For the covariance kernel, we use

$$\Gamma(s,t) = \sum_{j\geq 1} 2\theta_j \cos(j\pi s)\cos(j\pi t),$$

where $\theta_j = (|j - j_0| + 1)^{-2}$. When $j_0 = 1$, the two kernels are perfectly aligned, in the sense that they have the same sequence of eigenfunctions when ordered according to the eigenvalues. As j_0 increases, the level of mis-alignment also increases and we expect that the performances of the FPCA approach deteriorate with j_0. We set $n = 100, 200$ and $\sigma = 1, 3$, resulting in a total of four scenarios for each j_0. For values of j_0, we use $j_0 \in \{1, 2, 3, 4, 5\}$. For the FPCA approach, the tuning parameter is the truncation point which we consider in the range from 2 to 25. For the RKHS approach, the tuning parameter is λ and we consider $\lambda \in \exp\{-20, -19, \ldots, 0\}$. In the simulations, we assume the true sdr dimension 2 is known. The experiment for each scenario was repeated 100 times. In all situations the number of slices is set to be 10.

In this simulation, the tuning parameters are chosen to yield the smallest error to reflect the best achievable performance for both methods. Let P and $\widehat{P}$ be the orthogonal projection operators onto the true sdr space and the estimated sdr space respectively, the error is measured by the operator norm of $P - \widehat{P}$, denoted by $\|P - \widehat{P}\|_{op}$ with smaller values indicating better estimation performance. This distance is used in some previous works on sufficient dimension reduction such as Zhu et al. (2010). By Theorem I.5.5 of Stewart (1990), $\|P - \hat{P}\|_{op}$ is equal to the sine of the largest canonical

angle between the true and the estimated sufficient dimension reduction spaces. We also tried using the prediction risk as used in the theoretical analysis in the previous section and the results are similar and thus not reported.

Simulation results are summarized in Figure 1, which shows the error for both methods. Each panel corresponds to a pair of values of (n, σ), and the curves show the averaged error over 100 replications for both methods as j_0 increases (red curve for the FPCA approach and black curve for the RKHS approach). The vertical bar shows ± 2 standard errors computed from the 100 replications.

It is clearly seen that the performance of the RKHS approach is similar to that of the FPCA approach for $j_0 = 1$. As j_0 increases, the performance of the FPCA approach becomes much worse, while the errors for the RKHS approach remain at the same level. The difference in performance between these two methods generally increases with j_0.

In the second set of simulations, we investigate the case eigenfunctions of covariance and reproducing kernels are totally different. The data are generated from $Y_i = \langle \beta_1, X_i \rangle^3 + \langle \beta_2, X_i \rangle + \epsilon_i$, $\beta_1(t) = \sin(\pi t + 1), \beta_2(t) = \cos(\pi t + 1)$, $\epsilon \sim N(0, 0.5^2)$. X is generated as Brownian Motion with starting point randomly generated from a standard normal distribution.

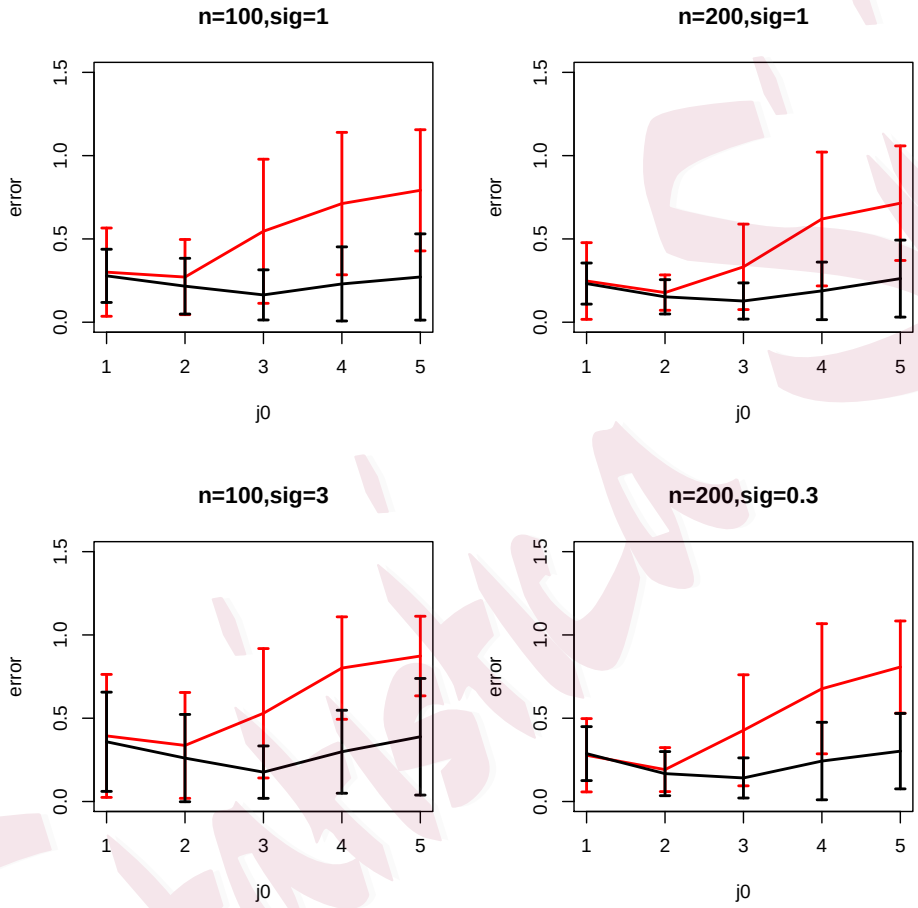

Figure 1: Error for both the FPCA method (red curve) and the RKHS method (black curve) for the first simulation example using the optimal tuning parameters.

For the RKHS approach, we set $\mathcal{H}_K$ to be the second-order Sobolev space W_2 as defined on page 7 of Wahba (1990) with the reproducing kernel given by $K(s,t) = 1 + st + \int_0^1 (t-u)_+(s-u)_+ du$. We set the sample sizes to be $n = 50, 100, 150, 200$. Simulation results are shown in Figure 2. We see that again the RKHS approach is better than PCA-based approach.

In general, selection of tuning parameter λ is a difficult task. When the ultimate goal is prediction, we can use cross-validation to select λ. More specifically, since we estimate two indices, two-dimensional Gaussian process regression is fitted (with the tgp package (Gramacy; 2007) in R) and 10-fold cross-validation is used to choose λ. The results are shown as the green curve in Figure 2. We see that cross-validation does a reasonably good job and the errors are close to the errors using the optimal λ.

4.2 Real data

We now turn to the prediction performance of the proposed method on a real dataset.

Canadian weather data. The daily weather data consists of daily temperature and precipitation measurements recorded in 35 Canadian weather stations. Each observation consists of functional data observed on an equally-spaced grid of 365 points. We treat the temperature as the

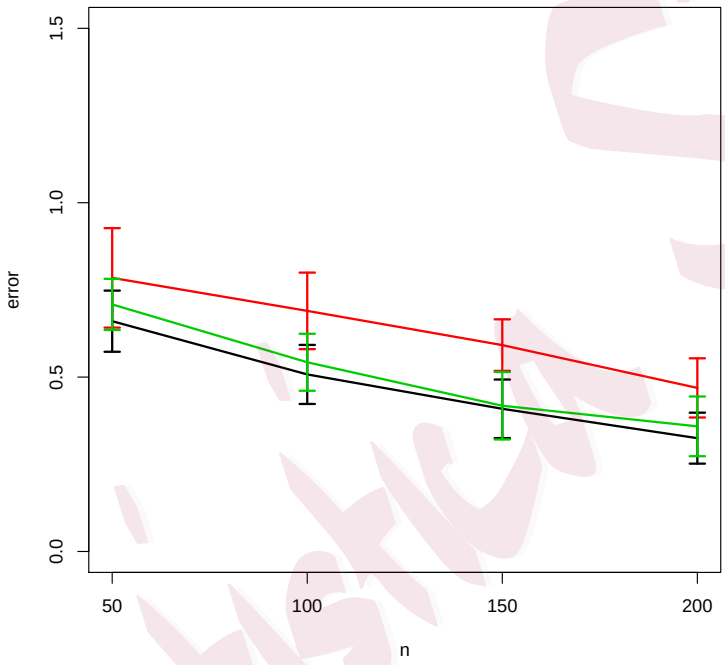

Figure 2: Error for both the FPCA method (red curve) and the RKHS method (black curve) for the second simulation example using the optimal tuning parameters. The green curve shows the results using λ selected by cross-validation.

independent variable and the goal is to predict the corresponding annual precipitation amount given the temperature measurements. As is previously done, we set the dependent variable to be the log-transformed precipitation. First, the number of indices need to be selected. For this we use the adaptive Neyman test proposed in (Li and Hsing; 2010), which is used for FSIR based on FPCA. Briefly, for any truncation level k, to test $H_0 : M \leq M_0$ vs. $H_a : M > M_0$, the test statistic is given by the sum of eigenvalues of an estimator of $Var(E[X|Y])$ except for the M_0 largest eigenvalues. Intuitively, this sum should be small if the null hypothesis $M \leq M_0$ is true. To remove the effect of the choice of k, the idea of the adaptive Neyman test is to standardize the different test statistics for different k and take the maximum. Asymptotic distribution of the test statistic is established in Li and Hsing (2010), allowing one to sequentially consider $M_0 = 0, 1, 2, \ldots$ and stops when one fails to reject the null. At significant level 0.05, the number of indices selected is 4 for the data set. Four eigenfunctions are then extracted in both the FPCA-based and the RKHS-based approaches. Given the periodic nature of the data, we set $\mathcal{H}_K = \mathcal{W}_2^{per}$, the second-order Sobolev space of periodic functions on $[0, 1]$. The reproducing kernel is given by $K(s,t) = 1 + \sum_{j\geq 1} \frac{2}{(2\pi j)^4} \cos(2\pi j(s-t))$. After the estimation of the 4 eigenfunctions, a four-dimensional Gaussian process

regression is fitted (with the tgp package (Gramacy; 2007) in R). We use leave-one-out cross-validation (CV) to determine the best tuning parameters to use for both methods. The average mean squared leave-one-out CV error for the FPCA-based approach is 0.178 while the error is 0.138 for the RKHS-based approach, with standard deviations 0.037 and 0.022, respectively. Furthermore, we can use the distance correlation to quantify the dependence between $\langle \beta, X \rangle$ and Y, which is a measure that characterizes independence, taking values in $[0, 1]$, and is zero if and only if the two random variables are independent. The distance correlations between $\langle \beta_j, X \rangle$ and Y, $j = 1, \ldots, 4$ are reported in Table 1 and show that the correlations for the RKHS based approach are larger, suggesting better performance. The four estimated index function $\beta_1, \ldots, \beta_4$ are shown in Figure 3 based on the proposed RKHS approach. Based on the shapes, we see that β_1 and β_3 focus on the contrast between temperature for the first half and the second half of a year, while β_2 concentrates on the summer months. β_4 has a periodic nature taking larger values in both very hot and very cold months.

5. Conclusion

In this paper, we established the minimax rate of convergence for estimation in functional sliced inverse regression in the general setting where the

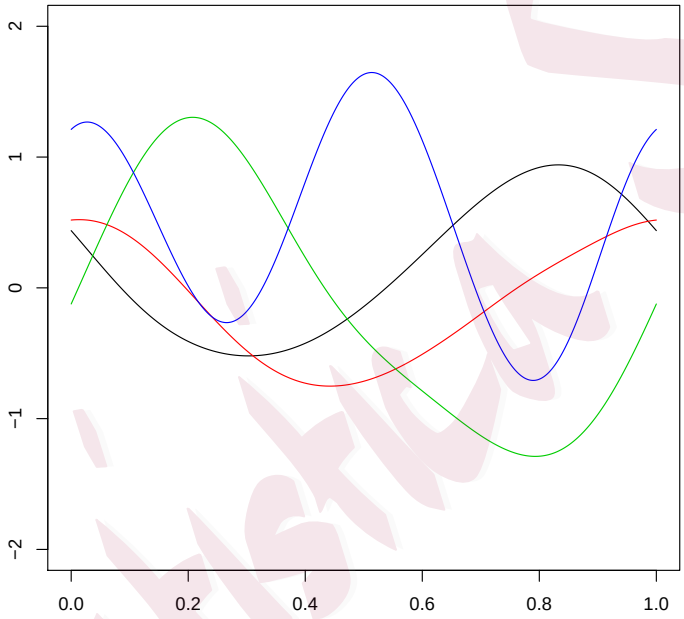

Figure 3: The estimated $\beta_1, \ldots, \beta_4$ based on the RKHS approach. The first to the fourth function is shown in black, red, green, blue, respectively.

Table 1: The estimated distance correlations for the estimated indices. The numbers in the brackets are the standard errors, which is computable from the multiple folds from cross-validation peroformed.

	β_1	β_2	β_3	β_4
FPCA	0.821(0.040)	0.481(0.079)	0.459(0.055)	0.364(0.055)
RKHS	0.869(0.044)	0.752(0.108)	0.654(0.025)	0.594(0.100)

covariance kernel Γ and the reproducing kernel K are not aligned, and also under general assumption on the decay rate of the eigenvalues of operator $T = K^{1/2}\Gamma K^{1/2}$. Our simulations show that as the degree of alignment of the two kernels decreases, the RKHS estimator can significantly outperform the estimator based on FPCA. The application on the weather data further demonstrates that the RKHS estimator can have better prediction accuracy.

We compared the results mainly with the method of Ferré and Yao (2003) which used the slicing estimator for the conditional expectation $E[X|Y]$. Ferré and Yao (2005) proposed to use the kernel estimator to estimate $E[X|Y]$. For the RKHS approach proposed here, we could also develop our methodology use the kernel estimator, parallel to Ferré and

Yao (2005). It is interesting to see the performance of the kernel estimator based approach which we leave as a future work.

The choice of smoothing parameter λ is generally hard. Thus in most of the simulations we choose the parameter that results in the smallest error. This is fine if the purpose is to get the best achievable performance in the simulations. This difficulty is not specific to the proposed method, and similar difficulty exists FPCA-based approach of Ferré and Yao (2003) since one needs to choose the truncation level k. On the other hand, if the prediction is the ultimate goal, one can use cross-validation to choose the parameter, as is done for both the simulation and the real data.

Given the well-known problem that FSIR sometimes cannot cover the entire sdr, it is natural to consider functional version of other sufficient dimension reduction approaches such as sliced average variance estimation (Cook and Weisberg; 1991) or directional regression (Li and Wang; 2007). We expect the method and theory developed in this paper can be extended to these estimation methods. However, given the more complicated form of these estimators, it may be challenging to demonstrate the convergence rate. Also, FSIR is not posed in an optimization framework, unlike the linear model (2.1). In the literature, optimization approach was sometimes used for sparse sufficient dimension reduction (Li; 2007; Chen et al.; 2010;

Lin et al.; 2016) and it might be more natural and interesting to extend the RKHS framework to these models. For example, by equation (7) of Lin et al. (2016), if one defines P to be the $n\times H$ matrix with entries $P_{ih} = I\{Y_i = y_h\}$, $\widehat{\eta}$ be the eigenfunction of $\widehat{Var}(E[X|Y])$ associated with its largest eigenvalue $\widehat{\mu}$ and $\widetilde{Y}_i = (H/(n\widehat{\mu}))\sum_{i',h} P_{ih}P_{i'h}\langle X_{i'}, \widehat{\eta}\rangle$, we can formulate the penalized functional as $\min_{\boldsymbol{\beta}\in\mathcal{H}_K} \sum_i(\widetilde{Y}_i - \int X_i\beta)^2 + n\lambda\|\beta\|^2_{\mathcal{H}_K}$, in a form the same as in (2.1). However, $\widetilde{Y}_i$ are no longer i.i.d. and study of the property of this estimator is challenging. We leave these for future investigations.

Supplementary Materials

Online supplementary materials contain the proofs of all technical results.

Acknowledgements The authors sincerely thank the Editor Professor Hsin-Cheng Huang, an Associate Editor, and two reviewers for their insightful comments that has led to significant improvements of the manuscript. The research of Heng Lian is supported by City University of Hong Kong Start-up Grant 7200521.

References

Birman, M. Š. and Solomjak, M. (1967). Piecewise-polynomial approximations of functions of the classes W_p^α, *Sbornik: Mathematics* **2**(3): 295–317.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression, *Annals of Statistics* **34**(5): 2159–2179.

Cai, T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression, *Journal of the American Statistical Association* **107**(499): 1201–1216.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model, *Statistics & Probability Letters* **45**(1): 11–22.

Cardot, H., Mas, A. and Sarda, P. (2007). Clt in functional linear regression models, *Probability Theory and Related Fields* **138**(3): 325–361.

Chen, X., Zou, C. and Cook, R. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection, *Annals of Statistics* **38**: 3696–3723.

Cook, R. and Lee, H. (1999). Dimension reduction in binary response regression, *Journal of the American Statistical Association* **94**(448): 1187–1200.

Cook, R. and Ni, L. (2005). Sufficient dimension reduction via inverse regression, *Journal of the American Statistical Association* **100**(470): 410–428.

Cook, R. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment, *Journal of the American Statistical Association* **86**(414): 328–332.

Duan, N. and Li, K. (1991). Slicing regression: a link-free regression method, *The Annals of Statistics* **19**(2): 505–530.

Ferré, L. and Yao, A. (2003). Functional sliced inverse regression analysis, *Statistics* **37**(6): 475–488.

Ferré, L. and Yao, A. (2005). Smoothed functional inverse regression, *Statistica Sinica* **15**(3): 665–683.

Gramacy, R. (2007). tgp: an r package for bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models, *Journal of Statistical Software* **19**(9).

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression, *Annals of Statistics* **35**(1): 70–91.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction, *Journal of the American Statistical Association* **102**(479): 997–1008.

Li, K. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association* **86**(414): 316–327.

Li, L. (2007). Sparse sufficient dimension reduction, *Biometrika* **94**(3): 603–613.

Li, Y. and Hsing, T. (2010). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data, *The Annals of Statistics* **38**(5): 3028–3062.

Lin, Q., Zhao, Z. and Liu, J. S. (2016). Sparse sliced inverse regression for high dimensional

data, *arXiv preprint arXiv:1611.06655* .

Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming, *Journal of Machine Learning Research* **13**: 389–427.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, MIT press Cambridge.

Stewart, G. W. (1990). *Matrix perturbation theory*, Academic Press, Boston.

Wahba, G. (1990). *Spline models for observational data*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Yao, F., Lei, E. and Wu, Y. (2015). Effective dimension reduction for sparse functional data, *Biometrika* **102**(2): 421–437.

Yao, F., Mueller, H. G. and Wang, J. L. (2005). Functional linear regression analysis for longitudinal data, *Annals of Statistics* **33**(6): 2873–2903.

Yin, X. and Cook, R. (2002). Dimension reduction for the conditional kth moment in regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(2): 159–175.

Zhu, L. and Fang, K. (1996). Asymptotics for kernel estimate of sliced inverse regression, *The Annals of Statistics* **24**(3): 1053–1068.

Zhu, L., Wang, T., Zhu, L. and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation, *Biometrika* **97**(2): 295–304.

College of Economics, Jinan University, Guangzhou, 510632, China

E-mail: wanggc023@amss.ac.cn

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong

E-mail: henglian@cityu.edu.hk