Statistics and Its Interface Volume 6 (2013) 187–196

Dimension reduction in functional regression using mixed data canonical correlation analysis

Guochang Wang, Nan Lin and Baoxue Zhang*

We propose a new dimension reduction method, mixed data canonical correlation (MDCANCOR), for functional regression with a scalar response and a functional predictor. MDCANCOR achieves dimension reduction using the canonical correlation analysis between the functional predictor and a set of B-spline basis functions that represent the transformed response space. And we propose a modified version of BIC to determine the dimensionality of the effective dimension reduction (EDR) space. This criterion is generally applicable to dimension reduction problems in functional regression. Asymptotically, we prove that MDCANCOR consistently estimates the directions when the dimensionality of the EDR space is given, and the modified BIC consistently estimates the dimensionality of the EDR space. Both simulation and real data examples show that the MDCANCOR method performs similarly as the regularized functional sliced inverse regression and better than other existing dimension reduction methods.

AMS 2000 subject classifications: Primary 62G05, 62G08; secondary 62G20.
Keywords and phrases: Dimension reduction, Effective dimension reduction, Functional regression, Mixed-data canonical correlation, splines.

1. INTRODUCTION

Functional data analysis (FDA) is widely used for data in the form of curves or functions. In this article, we focus on functional regression with a scalar response and a functional predictor. The basic approach to analyzing such data is functional linear regression [11, 17, 22, 28], while more flexible functional regression models can be found in [2, 5, 16, 27] and many others. A thorough discussion on this issue is given in [24, Chapter 15], and more comprehensive reviews of functional regression models can be found in Ramsay and Silverman [23, 24] and Ferraty and Vieu [10].

Let $L^2([a,b])$ be the space of squared integrable functions supported on $[a,b]$ with inner product $\langle f,g\rangle = \int_a^b f(t)g(t)dt$, and the associated norm be $\|f\| = \sqrt{\langle f,f\rangle}$. The functional linear regression model provides a natural extension of linear regression to the functional domain by assuming the relationship between a scalar response y and a functional predictor $x(t) \in L^2([a,b])$ as

$$y = \langle x,\beta\rangle + \varepsilon, \tag{1}$$

where ε is a random error term satisfying $E(\varepsilon) = 0$, $E(\varepsilon^2) < \infty$ and $E(x\varepsilon) = 0$, and $\beta \in L^2([a,b])$ is the functional regression coefficient. While the functional linear regression model (1) has gained great popularity for its simplicity, the linear form limits its application to data with more complicated structures. On the other hand, more flexible functional regression models often encounter the challenge of 'curse of dimensionality' as functional predictors are essentially of infinite dimensions. Researchers have developed methods based on functional principal component analysis to overcome the high-dimensional nature of the functional regression problem [18, 27]. However, a limitation of this type of method is that the reduced dimensions are derived regardless of the response variable and hence may not be relevant to the regression problem. Therefore, we consider regression-based dimension reduction and use a set of inner products to represent the reduced dimensions. More specifically, we focus on the following model

$$y = f(\langle\beta_1,x\rangle,\ldots,\langle\beta_K,x\rangle,\varepsilon), \tag{2}$$

where $\beta_1,\ldots,\beta_K$ are K linear independent functions in $L^2([a,b])$ that span a subspace E_K, and f is an unknown link function from $\mathbb{R}^{K+1}$ to $\mathbb{R}$. In dimension reduction literature, the space E_K is usually called the effective dimension reduction (EDR) space, and $\beta_1,\ldots,\beta_K$ are the EDR directions. A number of methods have been proposed for finding β_i's in (2) following the idea of sliced inverse regression (SIR) [21] for multivariate predictors, including functional SIR (FSIR) [7], functional inverse regression (FIR) [8], wavelet smoothing (WS) [1], and regularized FSIR (RFSIR) [9]. The aforementioned methods are all able to consistently estimate the EDR directions in that they guarantee that the estimated directions are contained in the EDR space. However, this property depends critically on a linearity condition on the predictor variable. When this condition fails, the estimated directions may no longer be consistent and become hard to interpret.

In this paper, we take a different approach to reducing the dimensionality of the functional predictor. Fung et al.

*Corresponding author.

[12] proposed CANCOR as an alternative to SIR for dimension reduction of multivariate predictors. The CANCOR method is based on a B-spline representation of the inverse link function, and estimates the EDR space as the span of the canonical variates from the canonical correlation analysis of the B-spline basis functions and the multivariate predictors. We propose the mixed-data canonical correlation (MDCANCOR) method as an extension of CANCOR to functional regression. MDCANCOR finds the EDR space using canonical variates based on the functional predictor x and the set of B-spline functions that represent the response variable y transformed by the inverse link function. MDCANCOR shares the same consistency property as other functional SIR-based approaches when the linearity condition on the predictor holds, and otherwise still gives the projections of the functional predictor x that are best correlated with the transformed response in terms of a penalized correlation. MDCANCOR is also computationally simpler than functional SIR-based methods.

A critical issue in dimension reduction methods is determining the dimensionality. For multivariate predictors, Li [21] developed a sequential chi-square test for deciding the dimensionality in the SIR method. This motivated a similar application in the functional case [1]. However, the authors did not provide any theory to show that the null distribution is still chi-square under the functional setup. In addition, Ferré and Yao [8] proposed a trace criterion, but this method requires an exhaustive search over a huge number of enumerations and hence computationally prohibitive in most applications. We treat the dimensionality as a modeling parameter and propose using a modified Bayesian information criterion (BIC) to determine it. Our proposal is motivated by Zhu et al. [31] who proposed a generic BIC-type procedure applicable to many multivariate dimension reduction methods. However, it is difficult to choose the penalty constant C_n in their criterion. Borrowing the ideal of the BIC, we propose another modified BIC based on the eigenvalues of the MDCANCOR operator, and we will illustrate the performance of the proposed MBIC by simulations.

In one of our earlier studies, Wang et al. [25] proposed FLIRST as a special application of MDCANCOR to the transformation model for functional linear regression

$$h(y) = \langle x, \beta_1 \rangle + \varepsilon, \tag{3}$$

where $h(\cdot)$ is a smooth function estimated by splines, and ε represents random noise independent of x. This special case essentially assumes the dimensionality of the EDR space is one, and hence the FLIRST method only uses the leading canonical variable in the MDCANCOR method. In this article, we investigate the usage and property of MDCANCOR for dimension reduction in a more general context. In particular, we derive in Section 3 the asymptotic property of the estimated covariance operator $\hat{\Gamma}_e$, where $\hat{\Gamma}_e$ is the estimate of $\Gamma_e = \mathrm{cov}(E(x|y))$. Under some mild assumptions, we prove that $\hat{\Gamma}_e$ converges to Γ_e at the rate of $n^{-1/2}$. In addition, in Section 4, we propose a modified BIC to determine the dimensionality of the EDR space and prove its consistency.

This paper is organized as follows. In Section 2, we introduce the MDCANCOR algorithm and the associated estimation procedure. Asymptotic properties are then derived in Section 3. In Section 4, we propose a modified BIC for determining the dimensionality in the MDCANCOR method and prove its consistency. Next, in Section 5, we demonstrate the merit of the MDCANCOR method by numerical examples, including two simulation studies and two real data examples. We then conclude the paper in Section 6. Technical proofs are given in the Appendix.

2. THE MDCANCOR METHOD

In this section, we propose the MDCANCOR method which is based on the canonical correlation between the functional predictor x and a B-spline basis representation of the response space.

2.1 Canonical directions

Following Fung et al. [12], without loss of generality, we suppose that the response variable y is supported on a bounded interval $[c, d]$. We first generate $\chi_n + l$ B-spline basis functions of order l with χ_n internal knots in $[c, d]$. Since the B-spline basis functions sum to 1 for each y, we need only use the first $\chi_n + l - 1$ basis functions of y, $\pi(y) = (B_1(y), \ldots, B_{\chi_n+l-1}(y))^T$. Suppose that the EDR space is K-dimensional. MDCANCOR solves an optimization problem that sequentially finds the canonical direction β_i's that give the maximum penalized correlation between the canonical variates $\langle x, \beta_i \rangle$ and a linear combination of the spline basis functions $b_i^T \pi(y)$ for $i = 1, \ldots, K$, where b_i are some unknown constant vectors.

Without loss of generality, suppose that the functional predictor x belongs to $L^2([a, b])$ with zero mean. Denote $\Gamma = E(x \otimes x)$ as the covariance operator of x, where $\otimes$ denotes the tensor product in $L^2([a, b])$. The tensor product for any $f,\ g \in L^2([a, b])$ is given by $(f \otimes g)(v) = \langle f, v \rangle g$ for any $v \in L^2([a, b])$. Denote by S the subspace of $L^2([a, b])$ containing functions with a squared integrable mth order derivative. We define the first canonical direction β_1 as the solution to the following optimization problem:

$$\max_{(\beta \in S, \mathbf{b} \in \mathbb{R}^{\chi_n + l - 1})} \mathrm{cov}(\langle x, \beta \rangle, \mathbf{b}^T \pi(y)), \tag{4}$$

subject to

$$\langle \Gamma\beta, \beta \rangle + \alpha_1 PEN_m(\beta) = \mathrm{var}(\mathbf{b}^T \pi(y)) = 1, \tag{5}$$

where $PEN_m(u) = \langle u^{(m)}, u^{(m)} \rangle$ defines the roughness (or smoothness) of the functional parameter β, and $u^{(m)}$ denotes the mth order derivative of u. Constraint (5) penalizes

the roughness of β, and uses a pre-specified positive smoothing parameter α_1 to control the degree of penalization. The penalization is necessary because the covariance operator Γ is otherwise empirically not estimable.

The rest of the canonical directions can be defined similarly. For identifiability, we require any subsequent canonical direction β_i $(i = 2, \ldots, K)$ to be uncorrelated with the preceding canonical directions $\beta_1, \ldots, \beta_{i-1}$. We then define the ith $(i \geq 2)$ canonical direction β_i as the solution to the following optimization problem:

$$\max_{(\beta \in S, \mathbf{b} \in \mathbb{R}^{\chi_n + l - 1})} \operatorname{cov}(\langle x, \beta \rangle, \mathbf{b}^T \pi(y)), \tag{6}$$

subject to

$$\langle \Gamma\beta, \beta \rangle + \alpha_i PEN_m(\beta) = \operatorname{var}(\mathbf{b}^T \pi(y)) = 1, \tag{7}$$

$$\langle \Gamma\beta, \beta_j \rangle + \alpha_i \langle \beta^{(m)}, \beta_j^{(m)} \rangle = 0, \quad j = 1, \ldots, i-1, \tag{8}$$

$$\operatorname{cov}(\mathbf{b}^T \pi(y), \mathbf{b}_j^T \pi(y)) = 0, \quad j = 1, \ldots, i-1. \tag{9}$$

Our definition then results in a set of canonical directions $(\beta_1, \ldots, \beta_K)$. In practice, the smoothing parameters α_i's in (5), (7) and (8) can be selected by cross-validation. However, our experience suggests that using a common value α for all i provides satisfactory performance and meanwhile reduces the numerical complexity. So, we will set all α_i being equal in the sequel.

Denote the penalty term in (5) and (7) by

$$Q_\alpha(f, f) = \langle \Gamma f, f \rangle + \alpha PEN_m(f). \tag{10}$$

Then through some simple algebra, it is easy to see that the optimization problems in (4) and (6) are equivalent to sequentially maximizing

$$\gamma_\alpha(\mathbf{b}_j, \beta_j) = \frac{(\operatorname{cov}(\langle x, \beta_j \rangle, \mathbf{b}_j^T \pi(y)))^2}{Q_\alpha(\beta_j)(\mathbf{b}_j^T \operatorname{var}(\pi(y)) \mathbf{b}_j)} \tag{11}$$

under the following set of orthogonal constraints, for $j = 1, \ldots, K$,

$$Q_\alpha(\beta_i, \beta_j) = \mathbf{b}_i^T \operatorname{var}(\pi(y) \mathbf{b}_j = \delta_{ij}, \quad \text{for all } i = 1, \ldots, j,$$

where $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$, and $\beta_j \in S$ and $\mathbf{b}_j \in \mathbb{R}^{\chi_n + l - 1}$.

Next, we give the key property of MDCANCOR. That is, the canonical directions defined by MDCANCOR are in the EDR space under a linearity condition.

Lemma 2.1. *Suppose that the following linearity condition holds.*

(a1) *For any $\beta \in L^2([a, b])$, there exists a vector $\mathbf{c} \in \mathbb{R}^K$ satisfying $E(\langle \beta, x \rangle | B) = \mathbf{c}^T B$, where $B = (\langle \beta_1, x \rangle, \ldots, \langle \beta_K, x \rangle)^T$.*

Then $E(x|y)$ belongs to the subspace spanned by $\Gamma\beta_1, \ldots, \Gamma\beta_K$.

Lemma 2.1 is Theorem 1 in Ferré and Villa [9], where readers are referred to find the proof.

Remark. The linearity condition (a1) is the same as condition 1 in Ferré and Yao [8] and is similar to the Condition 3.1 used in the multivariate case [21]. In the multivariate case, Hall and Li [13] proved that the linearity condition holds when the explanatory variable has a symmetric elliptical distribution or the dimension of the observed predictors is much larger than the dimension of the EDR space K. And it is often true that, even if the original predictor variables do not satisfy the linearity condition, some appropriate transformation will make it so. In the functional case, Ferré and Yao [7] proved a similar result that the linearity condition holds when the functional predictor is elliptically distributed, e.g. the Gaussian process.

If Condition (a1) does not hold, the directions found by MDCANCOR are not always guaranteed to be contained in the EDR space E_K but are still useful in identifying some main features of the regression model because the projections $\langle x, \hat{\beta}_j \rangle$ have the most significant penalized correlation with some transformation of the response variable y.

2.2 Estimation of the canonical directions

For real data, we estimate the canonical directions β_i's based on a sample version of the procedure described in Section 2.1. In reality, the functional predictor $x(t)$ is often observed only at a finite set of values of t. If the discrete observations are densely observed smooth data, for most theoretical and practical purposes, one can fit continuous curves to the discrete data and then treat the fitted curves as the true functional data; see, for example, [4, 29]. The case of sparsely observed functional data requires more careful treatment and can possibly be solved using the idea in Yao et al. [26]. In the present paper, we assume that the predictor curves are entirely observed in our theoretical development.

For a given random sample (x_i, y_i), $i = 1, \ldots, n$, we propose estimating the canonical directions β_i by replacing $\Gamma, \operatorname{cov}(\langle x, \beta \rangle)$ and $\operatorname{var}(\mathbf{b}^T \pi(y))$ in (11) by their sample versions, i.e. by maximizing the function

$$\hat{\gamma}_\alpha(\mathbf{b}_j, \beta_j) = \frac{(n^{-1} \sum_{i=1}^n \langle x_i - \bar{x}, \beta_j \rangle \mathbf{b}_j^T \pi^c(y_i))^2}{\hat{Q}_\alpha(\beta_j, \beta_j)(n^{-1} \mathbf{b}_j^T \sum_{i=1}^n (\pi^c(y_i))^T \pi^c(y_i) \mathbf{b}_j)}, \tag{12}$$

under orthogonal constraints, for $j = 1, \ldots, K$,

$$\hat{Q}_\alpha(\beta_i, \beta_j) = \mathbf{b}_i^T \sum_{i=1}^n (\pi^c(y_i))^T \pi^c(y_i) \mathbf{b}_j = \delta_{ij}, \quad \text{for all } i = 1, \ldots, j,$$

where $\pi^c(y_i) = \pi(y_i) - n^{-1} \sum_{j=1}^n \pi(y_j)$, $\bar{x} = n^{-1} \sum_{i=1}^n x_i$, $\hat{\Gamma} = n^{-1} \sum_{i=1}^n [x_i(s) - \bar{x}(s)] \otimes [x_i(t) - \bar{x}(t)]$ and $\hat{Q}_\alpha(f, f) = \langle \hat{\Gamma} f, f \rangle + \alpha PEN_m(f)$.

Next, we give the MDCANCOR procedure for estimating the canonical directions β_i's.

- Step 1. Choose a set of B-spline basis functions $\varphi_1(t), \ldots, \varphi_M(t)$ such that the functional parameters are represented as

$$\beta(t) = \sum_{i=1}^{M} a_i \varphi_i(t), \tag{13}$$

where a_i's are some unknown coefficients.
- Step 2. Let $\bar{x}(t) = n^{-1}\sum_{i=1}^{n} x_i(t)$ and $\bar{\pi} = \sum_{i=1}^{n} \pi(y_i)/n$. Centralize $(x_i, \pi(y_i))$ as $x_i^c(t) = x_i(t) - \bar{x}$ and $\pi^c(y_i) = \pi(y_i) - \bar{\pi}$.
- Step 3. Compute the $n \times M$ matrix C whose (i,j)th entry is $\langle x_i^c, B_j \rangle$ and the $M \times M$ matrix R with the (i,j)th entry being $\langle \varphi_i^{(m)}, \varphi_j^{(m)} \rangle$, and Π be the $n \times (\chi_n + l - 1)$ matrix which is the centered version of $(\pi(y_1), \ldots, \pi(y_n))^T$.
- Step 4. For a given smoothing parameter α, compute the eigenvalue-eigenvector decomposition of the matrix

$$\hat{\Psi} = (C^T C + n\alpha R)^{-1/2} \times C^T \Pi (\Pi^T \Pi)^{-1} \Pi^T C (C^T C + n\alpha R)^{-1/2}. \tag{14}$$

Let $\{\hat{a}_{ij}\}_{i=1,\ldots,K, j=1,\ldots,M}$ be the first K eigenvectors corresponding to the K largest eigenvalues of $\hat{\Psi}$. Then, we estimate the canonical directions $\{\beta_i(t)\}_{i=1,\ldots,K}$ by $\hat{\beta}_i(t) = \sum_{j=1}^{M} \hat{a}_{ij}\varphi_j(t)$, which span the EDR space E_K under Condition (a1).

3. ASYMPTOTIC PROPERTIES

In this section, we will show the consistency of the MDCANCOR method. First, we give some technical assumptions.

(a2) There is a positive constant δ_1 such that $E\|x\|^{4+\delta_1} < +\infty$;
(a3) For all $\alpha > 0$, $\rho_\alpha = \inf_{\|\beta\|=1, \beta \in S} Q_\alpha(\beta, \beta) > 0$;
(a4) $\lim_{n\to+\infty} \alpha = 0$, $\lim_{n\to+\infty} \sqrt{n}\alpha = \infty$;
(a5) There exist a unique maximizer $(\mathbf{b}_j^0, \beta_j^0)_{j=1,\ldots,K}$, $\mathbf{b}_j^0 \in \mathbb{R}^{\chi_n + l - 1}$ and $\beta_j^0 \in S$.
(a6) Assume that y has a probability density $f(y)$, and $f(y)$ is bounded away from 0 and infinity on $[a,b]$;
(a7) For every $t \in [a,b]$, $\zeta(v) = E(x(t)|y = v)$ is a function on $[c,d]$. Suppose $\zeta(v)$ have a $p'th$ derivative $\zeta^{p'}(v)$ such that

$$|\zeta^{p'}(v) - \zeta^{p'}(u)| \le C_1 |v - u|^{\tilde{p}}, \quad u, v \in [c,d],$$

where $C_1 > 0$ and $\tilde{p} \in (0,1]$, In what follows, we set $r = p' + \tilde{p}$;
(a8) $\chi_n = O_p(n^{\epsilon_1})$ and $e_n = O_p(n^{-\epsilon_2})$, where ϵ_1 and ϵ_2 are positive scalars such that $\epsilon_2/r + 1/(4r) < \epsilon_1 < 1/2 - 2\epsilon_2$;
(a9) The function $\varphi(v) = E(\|x\|^2 | y = v)$ is continuous and $\sqrt{n}E(\|\zeta(v)\|^2 I_{\{f(y) < e_n\}})$ tends to zero.

We first define the maximizer of the unregularized version of (11), i.e. $\alpha = 0$, as $(\mathbf{b}_j^0, \beta_j^0)$ and the maximum (or supreme) as λ_j. That is, $\lambda_j = \gamma_0(\mathbf{b}_j^0, \beta_j^0)$. The corresponding quantities for the sample version is then defined by $\lambda_j^n = \sup_{(\mathbf{b}_j \in \mathbb{R}^{\chi_n + l - 1}, \beta_j \in S)} \hat{\gamma}_\alpha(\mathbf{b}_j, \beta_j) = \hat{\gamma}_\alpha(\hat{\mathbf{b}}_j, \hat{\beta}_j)$.

The consistency here is defined through a mode of convergence in the functional space [19].

Definition. A sequence of functions $u_n(t)$ *converges in the* Γ*-norm* to $u(t)$ if $\mathrm{cor}^2(\langle x, u_n \rangle, \langle x, u \rangle) \to 1$ as $n \to \infty$, where x is the functional predictor.

Theorem 3.1. *Under the linearity condition* (a1) *and the assumptions* (a2)–(a5) *and* (a9)*, with probability converging to 1, the function* $\hat{\gamma}_\alpha(\mathbf{b}, \beta)$ *reaches its maximum on* $\mathbb{R}^{\chi_n + l - 1} \times S$ *when* n *grows to* $+\infty$*. Then, we have, as* $n \to +\infty$, $\gamma_0(\hat{\mathbf{b}}_j, \hat{\beta}_j) \xrightarrow{p} \lambda_j = \lim_{n\to+\infty} \lambda_j^n$*, and* $\hat{\beta}_j \to \beta_j$ *in the* Γ*-norm.*

Proof of the theorem is given in the Appendix A.1. Wang et al. [25] considered Theorem 3.1 by focusing on the leading canonical variate. Here, we present this theorem for all first K canonical variates and in addition, following [9], we prove that $\hat{\gamma}_\alpha(b, \beta)$ reaches its maximum on $\mathbb{R}^{\chi_n + l - 1} \times S$ with probability 1.

In the following, we will prove the consistency of

$$\hat{\Gamma}_e = \sum_{i=1}^{n} \frac{1}{n} \hat{\zeta}(y_i) \otimes \hat{\zeta}(y_i), \tag{15}$$

where $\Gamma_e = \mathrm{cov}(E(x|y))$ and $\hat{\zeta}(y_i)$ is the spline estimate of $\zeta(y_i) = E(x(t)|y = y_i)$. Denoting the Hilbert-Schmidt operator norm by $\|\cdot\|_{hs}$, and for any Hilbert-Schmidt operator A, take $\|A\|_{hs} = (\sum_j \|Ag_j\|^2)^{1/2}$, where (g_j) is any orthonormal basis in $L^2([a,b])$.

Theorem 3.2. *Under conditions* (2), (a2) *and* (a6)–(a9)*, we have* $\|\hat{\Gamma}_e - \Gamma_e\|_{hs} = O_P(n^{-1/2})$.

The proof of the Theorem 3.2 is given in the Appendix A.2.

Remark. Assumptions (a2)–(a5) are technical assumptions that ensure the existence and convergence of the estimated canonical directions $(\hat{\beta}_j)_{j=1,\ldots,K}$, and (a6)–(a9) are general assumptions that guarantee the convergence of $\|\hat{\Gamma}_e - \Gamma_e\|_{hs} = O_P(1/\sqrt{n})$. Assumption (a2) is essential to having $\hat{\Gamma}$ converge to Γ at the $\sqrt{n}$-rate. Assumption (a3), also used in [9], is due to regularization and controls the scaling of $\gamma_\alpha(\mathbf{b}, \beta)$. Assumption (a4) ensures that the denominator of $\hat{\gamma}_\alpha(\mathbf{b}, \beta)$ does not go too fast to zero. Assumption (a6)–(a9) are rather general assumptions which are also used in [8].

4. ESTIMATING THE DIMENSIONALITY OF THE EDR SPACE

In previous sections, we have assumed that the dimensionality of the EDR space K is known. In practice, it

is rarely given, though. MDCANCOR estimates the EDR space E_K by using the eigenvectors of Γ_e corresponding to the nonzero eigenvalues, where $\Gamma_e = \mathrm{cov}(E(x|y))$. Then, determining K is equivalent to estimating the number of nonzero eigenvalues of the operator Γ_e. For real data, Γ_e is estimated by $\hat{\Gamma}_e$ defined in (15).

Denote the eigenvalues of Γ_e by $\theta_1 \geq \cdots \geq \theta_M$, and the corresponding estimate is $\hat{\theta}_1 \geq \cdots \geq \hat{\theta}_M$. Zhu et al. [31] proposed a BIC-type procedure for determining the dimensionality of the EDR space for multivariate dimension reduction. The modified BIC proposed by Zhu et al. [31] is as follows, for $k = 0, \ldots, M-1$,

$$G(k) = \frac{n}{2} \sum_{i=1+\min(k,\tau)}^{M} (\log(\hat{\theta}_i + 1) - \hat{\theta}_i) - \frac{C_n k(2M-k+1)}{2},$$

where τ denote the number of $\hat{\theta}_i$'s that are greater than 0. This is a general method which can be applied to functional data by just treating the discretized observations of the functional predictor as a random vector. However, in their criterion, the choice of the penalty function C_n is difficult to choose in practice. To avoid the inconvenience of selecting C_n in the penalty term, we propose another modified BIC based on the eigenvalues of the MDCANCOR operator.

We define a modified BIC based on the estimated eigenvalues $\hat{\theta}_i$ as follows.

$$\mathrm{MBIC}(k) = n \frac{\sum_{i=1}^{k}(\hat{\theta}_i^2)}{\sum_{i=1}^{M}(\hat{\theta}_i^2)} - \frac{\log(n)k}{2M}. \tag{16}$$

The second term of MBIC is a penalty with k being the number of $\hat{\theta}_i$ needed to be estimated. Similar to BIC, we include the factor $\log(n)$ in the penalty. Then we estimate K by

$$\hat{K} = \underset{1 \leq k \leq M}{\mathrm{argmax}}\, \mathrm{MBIC}(k). \tag{17}$$

The next theorem shows that this criterion consistently estimates the dimensionality K.

Theorem 4.1. *Under the assumptions of Theorem 3.2, the estimated dimension $\hat{K}$ converges to K in probability.*

The proof of Theorem 4.1 is similar with the Theorem 2 of Zhu et al. [31] and hence omitted, and we refer the readers to Zhu et al. [31] for more details.

5. NUMERICAL STUDIES

5.1 Simulation study

In this section, we carry out simulations to study the performance of MDCANCOR. We compare our method with existing methods, including FSIR [7], FIR [8], WS [1] and RFSIR [9], in their prediction accuracy and estimated EDR dimensions.

To measure the predictive performance of different methods, we use the root mean squared prediction error, $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$, where $\hat{y}_j$ denotes the predicted value and y_j is the corresponding true value. In the simulation, we compute $\hat{y}_j$ through replacing β_j in the true model by its estimate $\hat{\beta}_j$, $j = 1, \ldots, K$. For real data, similar to Amato et al. [1], we compute $\hat{y}_j = \sum_{i=1}^{K}\langle x_j, \hat{\beta}_j\rangle$.

Another way to compare the performance of different methods is by the accuracy of estimating the EDR dimensions. We adopt the squared trace correlation coefficient [8] between $\langle\beta_1, x\rangle, \ldots, \langle\beta_K, x\rangle$ and $\langle\hat{\beta}_1, x\rangle, \ldots, \langle\hat{\beta}_K, x\rangle$. Empirically, we evaluate it by discretization on a set of points $t_1, \ldots, t_D$ on $[a, b]$, i.e.

$$R^2(\beta) = \frac{trace[\beta(\Sigma\beta)^T\hat{\beta}(\Sigma\hat{\beta})^T]}{K},$$

where $\beta = ((\beta_1(t_1), \ldots, \beta_1(t_D))^T, \ldots, (\beta_K(t_1), \ldots, \beta_K(t_D))^T)$, $\hat{\beta} = ((\hat{\beta}_1(t_1), \ldots, \hat{\beta}_1(t_D))^T, \ldots, (\hat{\beta}_K(t_1), \ldots, \hat{\beta}_K(t_D))^T)$ and Σ is the $D \times D$ matrix given by $\hat{\Gamma}(t_i, t_j)$, $i, j = 1, \ldots, D$. For a given data set, $t_1, \ldots, t_D$ are taken as those on which the functional predictor x_i is measured.

For the considered methods, we need to decide the following tuning parameters, 1) for MDCANCOR: the number of knots χ_n and the order of the B-spline basis function l for $\pi(y)$, the smoothing parameter α, the order of the derivative m, and the number of the basis function M in (13); 2) for FSIR and WS: the number of slices H; 3) for FIR: the bandwidth h; 4) for RFSIR: the number of slices H and a smoothing parameter λ.

Our experience suggested that the results are not sensitive to the choice of χ_n, l, m, M and H among the above tuning parameters. In all simulations, we set $\chi_n = 9$, $l = 3$, $m = 3$, $M = 15$ and $H = 15$. Other tuning parameters are chosen by 10-fold cross-validation. Suppose that the subjects are partitioned into 10 subsets $\{\omega_1, \ldots, \omega_{10}\}$. Taking α as an example, we select α as

$$\hat{\alpha} = \underset{\alpha>0}{\mathrm{argmin}} \sum_{i=1}^{10}\sum_{j\in\omega_i}(y_j - \hat{y}_j^{(-\omega_i)})^2, \tag{18}$$

where $\hat{y}_j^{(-\omega_i)}$ is the predicted value by using samples in $\{\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \omega_{10}\}$.

Example 5.1. Here we consider the following two models,

$$\begin{aligned} \text{Model A:}\quad & y_i = \sin\left(\tfrac{\pi}{2}\langle x_i, \beta_1\rangle\right) + \langle x_i, \beta_2\rangle + \varepsilon_i, \\ \text{Model B:}\quad & y_i = \sin\left(\tfrac{\pi}{2}\langle U_i, \beta_1\rangle\right) + \langle U_i, \beta_2\rangle + \varepsilon_i, \\ & \text{with } U_i(t_j) = x_i(t_j) + \tilde{\varepsilon}_{ij}, \end{aligned}$$

where $x_i(t)_{t\in[0,1]}$ is taken as the standard Brownian motion and ε_i is from the standard normal distribution and independent of $x_i(t)_{t\in[0,1]}$, and $\tilde{\varepsilon}_{ij}$ from $N(0, 0.25)$ is the measurement error. Mode A was previously considered in Ferré

*Table 1. Average (*standard error*) of RMSE and $R^2(B)$ for FSIR, FIR, WS, RFSIR and MDCANCOR for Example 5.1*

	Model	FSIR	FIR	WS	RFSIR	MDCANCOR
A	RMSE	1.522(*0.110*)	1.508(*0.107*)	1.493(*0.186*)	1.072(*0.096*)	1.063(*0.090*)
	$R^2(B)$	0.90(*0.021*)	0.92(*0.019*)	0.92(*0.068*)	0.94(*0.023*)	0.95(*0.025*)
B	RMSE	1.523(*0.111*)	1.510(*0.108*)	1.512(*0.187*)	1.089(*0.121*)	1.070(*0.100*)
	$R^2(B)$	0.89(*0.024*)	0.92(*0.021*)	0.91(*0.071*)	0.94(*0.023*)	0.94(*0.030*)

*Table 2. Average (*standard error*) of RMSE and $R^2(B)$ for FSIR, FIR, WS, RFSIR and MDCANCOR for Example 5.2*

	Model	FSIR	FIR	WS	RFSIR	MDCANCOR
A	RMES	1.385(*0.257*)	1.151(*0.146*)	1.184(*0.168*)	1.073(*0.087*)	1.067(*0.096*)
	$R^2(B)$	0.89(*0.021*)	0.90(*0.059*)	0.89(*0.116*)	0.94(*0.021*)	0.95(*0.023*)
B	RMES	1.387(*0.259*)	1.161(*0.154*)	1.192(*0.170*)	1.076(*0.0902*)	1.070(*0.101*)
	$R^2(B)$	0.89(*0.024*)	0.89(*0.066*)	0.88(*0.118*)	0.95(*0.022*)	0.95(*0.026*)

and Yao [8], and Model B is a contaminated version with measurement error. In these models, $\beta_1(t) = (2t-1)^3+1$ and $\beta_2(t) = \cos(\pi(2t-1))+1$ span the EDR space. To make the directions of the functional space identifiable, we consider the Γ-orthonormed β_1 and β_2 in calculating $R^2(\beta)$ and the RMSE. That is, we impose the constraints $\langle \Gamma\beta_i, \beta_j\rangle = \delta_{ij}$ for $i, j = 1, 2$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. In this study, we simulate 2,000 Monte Carlo samples each of size 400 from model A and model B of Example 5.1. and each random curve $x_i(t)$ is sampled at $p = 128$ equally spaced points in the interval $[0, 1]$. To evaluate the prediction performance, the sample was divided into two parts, a training sample of size 300 to estimate the EDR space, and a test sample of size 100 to compute the RMSE. In computing the RMSE and $R^2(\beta)$, we use the true dimensionality of the EDR space, i.e. we set $K = 2$ for all methods.

In Table 1, we observe that MDCANCOR outperforms other methods with the smallest RMSE and the largest trace correlation. RFSIR is nearly as good as MDCANCOR, whereas FIR, WS and FSIR have significantly larger RMSE and a smaller trace correlation. Among the last three, we also noticed that FIR and WS perform similarly and better than FSIR.

Example 5.2. In this simulation study, we consider the following models,

$$\begin{aligned} \text{Model A:} \quad & y_i = \left(\langle x_i, \beta_1\rangle\right)^2 + |\langle x_i, \beta_2\rangle| + \varepsilon_i, \\ \text{Model B:} \quad & y_i = \left(\langle U_i, \beta_1\rangle\right)^2 + |\langle U_i, \beta_2\rangle| + \varepsilon_i, \\ & \text{with } U_i(t_j) = x_i(t_j) + \tilde{\varepsilon}_{ij}, \end{aligned}$$

where $x_i(t)$ is from the standard Brownian motion on $t \in [0, 1]$ and ε_i is the standard normal random error that is independent of $x_i(t)$. The rest of the setups are the same as in Example 5.1 and tuning parameters are also decided in the same way. It is well known in multivariate dimension reduction, the SIR-based method have poor performance on the non-monotonic trends in the dependence of y on x [20]. Compared with Example 5.1, Example 5.2 is based on a more complex regression surface in which no simple monotonic function exists. In this sense, the functional parameters β_1 and β_2 in Example 5.2 are harder to estimate than in Example 5.1.

Table 2 reports the average RMSE and $R^2(\beta)$ over 2,000 simulations. Similarly as in Example 5.1, we observe that MDCANCOR and RFSIR have very close performance and are better than the rest of the methods.

For both Examples 5.1 and 5.2, results in Tables 1 and 2 show almost no difference in the results for Model A and Model B. This phenomenon indicates the impact of measurement error is minor in these examples and it agrees with the theory of Zhang and Chen [29]. From Tables 1 and 2, we also observe that MDCANCOR and RFSIR have similar performance and are better than other methods, among which FIR and WS are close and leave FSIR behind. Such patterns were also observed previously in Ferré and Villa [9] and Wang et al. [25].

Table 3 reports the frequencies of the EDR space dimensionality selected by the MBIC and the trace criterion $R(q)$ [8] in Examples 5.1 and 5.2. In both examples, it is seen that MBIC gives a higher frequency than the trace criterion in selecting the correct dimensionality $K = 2$. And both criteria perform better in Example 5.1 than in Example 5.2, which is likely because it is easier for SIR-based methods to identify models in Example 5.1. Further, the selection results under Model A and Model B are similar, which again shows that the impact of measurement error is quite minor.

5.2 Real data examples

In this subsection, we compare MDCANCOR with FSIR, FIR, WS and RFSIR on two real data examples, *Tornado data* and *South Dakota data*.

Tornado Data. The Tornado data, previously analyzed by Baíllo and Grané [3], are from the U.S. National Climatic Data Center website (www.ncdc.noaa.gov). The response variable y_i is the logarithm of the total number of

Table 3. Frequencies of selected model dimension by modified BIC and trace criterion $R(q)$ for Example 5.1 and Example 5.2 for the training sample

		Method	$K=0$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
Example 5.1	Model A	MBIC	0	258	1739	3	0	0
		$R(q)$	0	3	1142	855	0	0
	Model B	MBIC	0	264	1726	10	0	0
		$R(q)$	0	2	1133	865	0	0
Example 5.2	Model A	MBIC	0	208	1262	471	57	2
		$R(q)$	1	129	1086	751	21	12
	Model B	MBIC	0	189	1254	535	22	0
		$R(q)$	0	107	1059	790	44	0

Table 4. RMSE for (a) Tornados data and (b) South Dakota data

	FSIR	FIR	WS	RFSIR	MDCANCOR
(a)	0.5486	0.6994	0.7007	0.5121	0.4908
(b)	0.1829	0.1784	0.1856	0.1736	0.1702

tornados in each U.S. state ($i = 1, \ldots, 48$) during the period of 2000-2005. The predictor variable x_i is the monthly average temperature (in Fahrenheit) in state i in the same period of time, which contains 72 discrete observations. This is of interest, for instance, when assessing the possible consequence of an overall temperature increase due to climate change.

South Dakota data. This data set contains daily maximum temperatures (in Fahrenheit) and the total precipitation over the course of the year 2000 at $n = 80$ weather stations from South Dakota. The response variable y_i is the logarithm of the total precipitation in each of the stations during the same year, and the predictor x_i is measured by the 365 daily maximum temperatures. The aim is to predict the logarithm of the total precipitation y_i given the discrete observations $x_i(t_j)$.

In both examples, we choose the tuning parameters in the same way as in simulation studies, except that we use a smaller number of slices $H = 4$ because of the small sample size. For all methods, we use the modified BIC (16) to select the dimensionality of the EDR space.

In Table 4, we summarized the RMSE via a cross-validation procedure, i.e.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2},$$

where $\hat{y}_i^{(-i)}$ is the predicted value based on $(n-1)$ observations with the observation (x_i, y_i) omitted.

We select the tuning parameters by the 10-fold cross-validation for both of the real data analysis as in the simulation studies. For both examples, Table 4 shows that our MDCANCOR method performs the best with the smallest RMSE while RFSIR seems a close competitor. Meanwhile, the WS method is seen to give the largest RMSE in both examples. This likely due to the fact that the WS method results in a lower dimensionality K when the modified BIC is applied.

6. CONCLUSION

In this paper, we propose MDCANCOR for dimension reduction in functional regression. Our method uses a spline representation of the transformed response and then estimates the EDR space using the canonical correlation analysis between the functional predictor and the multivariate vectors constructed by the spline basis functions. Canonical directions estimated by MDCANCOR are shown to be consistent estimates of the EDR directions under the linearity condition (a1).

From Wang et al. [25], we know that MDCANCOR and RFSIR are estimating the same quantities and are consistent for functional directions in the EDR space under the linear condition (a1). However, the directions estimated by MDCANCOR are more directly interpretable outside the context of the EDR space. Compared with RFSIR, MDCANCOR is not only applicable to independent random samples. With some modified technical assumptions, it can also be extended to data with autoregressive Hilbertian functional predictor. Simulations show that MDCANCOR has a close performance with RFSIR and outperforms other existing dimension reduction methods for functional regression. Furthermore, we propose a new method to determine the dimension of the EDR space.

Regularization is essential to functional dimension reduction methods as the functional variable is infinitely dimensional. We have adopted a penalty of the bilinear form in (5), however, our theory should hold also for other regularization functionals satisfying the assumption (a3). For example, we

may use a penalty term similar to the one used in Ridge-PDA [14], which will lead the roughness penalty terms in (5) to $\langle(\Gamma+\lambda I)\beta,\beta\rangle = \mathrm{var}(\mathbf{b}^T\pi(y)) = 1$.

MDCANCOR achieves dimension reduction from a perspective different from SIR-based methods, such as FSIR, WS, FIR and RFSIR, and may lead to some flexible extensions. One direction is in robustifying the method. As MDCANCOR is based on the classical estimates of the first and second linear operator of the data, it can be sensitive to outliers. In the multivariate case, WCANCOR [30] is proposed as a robust version of CANCOR by downweighing the outlying observations. It is quite natural to achieve a robust functional dimension reduction method by extending MDCANCOR in a similar way. On the other hand, robustification of the SIR-based functional dimension reduction methods is not as straightforward as finding the weights can be very difficult. A second direction to extend MDCANCOR is to consider functional regression with multivariate responses, i.e. $\mathbf{y} \in R^p$. FSIR and RFSIR can be extended in theory to multivariate response without modification but are faced with the curse of dimensionality in practice. For example, if one slices each response variable into a fixed number of slices h, then the total number of slices h^p increases exponentially, and therefore a great amount of data will be required to fill in each slice when the dimension p is high. Such challenges will not occur in extending our MDCANCOR method, as canonical correlation analysis is originally designed for handling multivariate responses, and we may actually achieve simultaneous dimension reduction on the response side by finding a linear combination of the multivariate responses as a scalar surrogate. We will investigate these extensions in our future research.

APPENDIX A. PROOF OF THEOREMS

A.1 Proof of Theorem 3.1

The proof of Theorem 3.1 is similar to Theorem 2 of [9]. To prove Theorem 3.1, we need Theorem 1 in Leurgans et al. [19]. We present it as a lemma in the following. Let $\Gamma_e = E(E(x|y)\otimes E(x|y))$ be the covariance operator $E(x|y)$ and $\hat{\Gamma}_e = C^T\Pi/n(\Pi^T\Pi/n)^{-1}\Pi^T C/n$, where C and Π are defined in subsection 2.2.

Lemma A.1. *Let*

$$\delta^n = \max\{\|\hat{\Gamma}-\Gamma\|; \|\hat{\Gamma}_e-\Gamma_e\|\}.$$

If the sequence κ_n satisfies $\sqrt{n}\kappa_n \to +\infty$, we have $\kappa_n^{-1}\delta^n \xrightarrow{p} 0$.

Next, we will give the proof of Theorem 3.1.

Proof. (i) Existence. For $\alpha\in(0,1)$, we can easily get $Q_\alpha = (1-\alpha)\langle\Gamma.,.\rangle + \alpha Q_1$. Furthermore, by the positiveness of Γ, we have $\alpha^{-1}Q_\alpha(u,u) = (\alpha^{-1}-1)\langle\Gamma u,u\rangle + Q_1 > \rho_1$ for all u satisfying $\|u\|=1$. Then, by the Assumption (a3) we have $\sqrt{n}\rho_\alpha > \alpha\sqrt{n}\rho_1$ and

$$\sqrt{n}\rho_\alpha \to +\infty. \tag{19}$$

For notational convenience, we define $\triangle_1^n = \hat{\Gamma}-\Gamma$. By Lemma A.1, we have $\lim_{n\to+\infty} P(\{\omega\in\Xi : \|\triangle_1^n\| \le \rho_\alpha/2\}) = 1$, where Ξ denotes the probability space on which x and y are defined. Then, since

$$\{\omega\in\Xi : \|\triangle_1^n\| \le \rho_\alpha/2\} \subset \{\omega : \forall\vartheta\in S, \|\vartheta\|=1, \hat{Q}_\alpha(\vartheta,\vartheta)\ge\rho_\alpha/2>0\},$$

we get $\lim_{n\to+\infty} P(\{\omega : \forall\vartheta\in S, \|\vartheta\|=1, \hat{Q}_\alpha(\vartheta,\vartheta)\ge\rho_\alpha/2\}) = 1$, where $\hat{Q}_\alpha(\vartheta,\vartheta) = \langle\hat{\Gamma}\vartheta,\vartheta\rangle + \alpha PEN_m(\vartheta)$.

Denote $\overline{B}$ as the weak closure of $B = \{(\mathbf{b},\vartheta), \vartheta\in S, \mathbf{b}\in\mathbb{R}^{\chi_n+l-1} : \hat{Q}_\alpha(\vartheta,\vartheta)=1, n^{-1}\mathbf{b}^T\sum_{i=1}^n(\pi^c(y_i))^T\pi^c(y_i)\times\mathbf{b}=1\}$, and let ζ be the function defined on B by $\zeta(\mathbf{b},\vartheta) = n^{-1}\sum_{i=1}^n\langle x_i,\vartheta\rangle\mathbf{b}^T\pi(y_i)$. Let $\tilde{\zeta}$ be a uniformly continuous function defined on $\overline{B}$ for the weak topology, which is the extension of the function ζ. Finally, provided that $\hat{Q}_\alpha(\vartheta,\vartheta)\ge\rho_\alpha/2$, $\tilde{\zeta}$ reaches its maximum on weak compact $\overline{B}$, which concludes the proof of the existence of $(\hat{\mathbf{b}}_j,\hat{\beta}_j)_{j=1,\dots,K}$.

(ii) Consistency. In the following, we consider $\tilde{\omega}\in\Xi$ such that $\tilde{\omega}\in\{\omega\in\Xi$: under orthogonal constraints, $\hat{\gamma}_\alpha(\mathbf{b},\beta)$ has maximum on $\mathbb{R}^{\chi_n+l-1}\times S$ and reaches it$\}$. Let λ_j^n be these maximum (i.e. $\lambda_j^n = \hat{\gamma}_\alpha(\hat{\mathbf{b}}_j,\hat{\beta}_j)$, under orthogonal constraints) and λ_j^α be the jth maximum of $\gamma_\alpha(\mathbf{b},\beta)$ on $\mathbb{R}^{\chi_n+l-1}\times S$, and the corresponding maximizers be $(\mathbf{b}_j^\alpha,\beta_j^\alpha)$. Here, λ_j^α is well defined according to assumption (a3).

We first show $\gamma_0(\hat{\mathbf{b}}_j,\hat{\beta}_j)\xrightarrow{p}\lambda_j$ for $j=1,\dots,K$. For given $\mathbf{b}$ and β, we have

$$\frac{\gamma_\alpha(\mathbf{b},\beta)}{\gamma_0(\mathbf{b},\beta)} = \frac{\langle\Gamma\beta,\beta\rangle}{\langle\Gamma\beta,\beta\rangle+\alpha PEN_m(\beta)} \le 1, \tag{20}$$

because Γ and $PEN_m(.)$ are non-negative definite. Thus $\lambda_j^\alpha = \gamma_\alpha(\mathbf{b}_j^\alpha,\beta_j^\alpha) \le \gamma_0(\mathbf{b}_j^\alpha,\beta_j^\alpha) \le \lambda_j$. It follows from the equality in expression (20) that, for any fixed $\mathbf{b}$ and β, $\gamma_\alpha(\mathbf{b},\beta)\to\gamma_0(\mathbf{b},\beta)$ as $\alpha\to0$, and so we have $\lambda_j\ge\lambda_j^\alpha\ge\gamma_\alpha(\mathbf{b}_j^0,\beta_j^0)\to\gamma_0(\mathbf{b}_j^0,\beta_j^0)=\lambda_j$. So we obtain

$$\lambda_j^\alpha \to \lambda_j. \tag{21}$$

Furthermore, by the law of large numbers, it is easy to show that $\sup_{(\mathbf{b}\in\Re^{\chi_n+l-1},\beta\in S)}|\hat{\gamma}_\alpha(\mathbf{b},\beta)-\gamma_\alpha(\mathbf{b},\beta)|\xrightarrow{p}0$. Then under orthogonal constrains (10), we can show that

$$|\lambda_j^n-\lambda_j^\alpha|\xrightarrow{p}0. \tag{22}$$

Finally, by combining (21) and (22), we can get that

$$\lambda_j^n\xrightarrow{p}\lambda_j. \tag{23}$$

From (23), we then have

$$\lambda_j \geq \gamma_0(\hat{\mathbf{b}}_j, \hat{\beta}_j) \geq \gamma_\alpha(\hat{\mathbf{b}}_j, \hat{\beta}_j) \xrightarrow{p} \hat{\gamma}_\alpha(\hat{\mathbf{b}}_j, \hat{\beta}_j) = \lambda_j^n \xrightarrow{p} \lambda_j,$$

which leads to

$$\gamma_0(\hat{\mathbf{b}}_j, \hat{\beta}_j) \xrightarrow{p} \lambda_j = \gamma_0(\mathbf{b}_j^0, \beta_j^0). \tag{24}$$

Second, we assume without loss of generality (multiplying β_j, $\hat{\beta}_j$ and $\mathbf{b}_j$ by suitable constants if necessary) that $\langle \Gamma\beta_j, \beta_j\rangle = \langle \Gamma\hat{\beta}_j, \beta_j\rangle = \mathbf{b}_j^T \mathrm{var}(\pi(y))\mathbf{b}_j = 1$. It is apparent that this rescaling will not affect any of the correlations in this theorem. Now decompose $\hat{\beta}_j = \beta_j + \tilde{\beta}_j$, where $\langle \Gamma\beta_j, \tilde{\beta}_j\rangle = 0$ and similarly $\hat{\mathbf{b}}_j = \mathbf{b}_j + \tilde{\mathbf{b}}_j$, where $\mathbf{b}_j^T \mathrm{var}(\pi(y))\tilde{\mathbf{b}}_j = 0$.

Next, we will show $\varsigma_j^2 = \langle \Gamma\tilde{\beta}_j, \tilde{\beta}_j\rangle \to 0$. First, because $\hat{Q}_\alpha(\hat{\beta}_j) = 1$ and $\alpha PEN_m(\hat{\beta}_j) \geq 0$, we have $\langle \hat{\Gamma}\hat{\beta}_j, \hat{\beta}_j\rangle = \hat{Q}_\alpha(\hat{\beta}_j) - \alpha PEN_m(\hat{\beta}_j) \leq 1$. Second, we have

$$\langle \Gamma\hat{\beta}_j, \hat{\beta}_j\rangle = \langle \Gamma\beta_j, \beta_j\rangle + \langle \Gamma\tilde{\beta}_j, \tilde{\beta}_j\rangle = 1 + \langle \Gamma\tilde{\beta}_j, \tilde{\beta}_j\rangle \geq 1. \tag{25}$$

Then it follows that $1 \geq \langle \hat{\Gamma}\hat{\beta}_j, \hat{\beta}_j\rangle \xrightarrow{p} \langle \Gamma\hat{\beta}_j, \hat{\beta}_j\rangle \geq 1$, i.e. $\langle \Gamma\hat{\beta}_j, \hat{\beta}_j\rangle = 1$. Hence, it is obvious that $\varsigma_j^2 \to 0$.

Consequently, the convergence in Γ-norm follows from $\mathrm{cor}^2(\langle x, \hat{\beta}_j\rangle, \langle x, \beta_j\rangle) = \langle \Gamma\beta_j, \hat{\beta}_j\rangle^2 / (\langle \Gamma\hat{\beta}_j, \hat{\beta}_j\rangle \langle \Gamma\beta_j, \beta_j\rangle) = (1 + \varsigma_j^2)^{-1} \to 1$. And this completes the proof of Theorem 3.1. □

A.2 Proof of Theorem 3.2

Proof. To prove Theorem 3.2, we recall first that the estimator $\hat{\zeta}(v)$ of $\zeta(v)$ is obtained by spline smoothing of $x(t)$ with design points y_i's, $i = 1, \ldots, n$. By the equivalent kernel method for least squares spline regression [15], similar to a Naradaya-Watson estimator, we define the estimator $\hat{\zeta}(v)$ as the ratio of two estimators, the numerator $\hat{h}$ being an estimate of $h = \zeta(v)f$ and the denominator $\hat{f}(y)$ being a spline estimate of the marginal density of y. To prevent the sensitivity to small values of f, we threshold f by a sequence $\{e_n\}$ converging to zero. Let $f_{e_n} = \max(f, e_n)$ and $\hat{f}_{e_n} = \max(\hat{f}, e_n)$. We define $\hat{\zeta}_n(v) = \hat{h}/\hat{f}_{e_n}$, and it retains the same asymptotic properties as $\hat{\zeta}(v)$ following the results in [6]. For simplicity, in this proof, we still denote $\hat{\Gamma}_e = \frac{1}{n}\sum_{i=1}^n \hat{\zeta}_n(y_i) \otimes \hat{\zeta}_n(y_i)$.

Let $\zeta_n(y) = h(y)/f_{e_n}(y)$, $\bar{\Gamma}_e = \frac{1}{n}\sum_{i=1}^n \zeta_n(y_i) \otimes \zeta_n(y_i)$ and $\tilde{\Gamma}_e = \frac{1}{n}\sum_{i=1}^n \zeta(y_i) \otimes \zeta(y_i)$. Then

$$\hat{\Gamma}_e - \Gamma_e = (\hat{\Gamma}_e - \bar{\Gamma}_e) + (\bar{\Gamma}_e - \tilde{\Gamma}_e) + (\tilde{\Gamma}_e - \Gamma_e). \tag{26}$$

By the weak law of large numbers, we know that the last term in (26) satisfies

$$\tilde{\Gamma}_e - \Gamma_e = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{27}$$

Through the same deduction as in Ferré and Yao [8], we can get

$$\hat{\Gamma}_e - \bar{\Gamma}_e = o_p\left(\frac{1}{\sqrt{n}}\right), \quad \text{and} \quad \bar{\Gamma}_e - \tilde{\Gamma}_e = o_p\left(\frac{1}{\sqrt{n}}\right). \tag{28}$$

Then, combining (26), (27) and (28), we have $\hat{\Gamma}_e - \Gamma_e = O_p(\frac{1}{\sqrt{n}})$ and the proof is complete. □

ACKNOWLEDGEMENTS

We would like to thank the referees and the editor for careful reading of a previous version of the manuscript that allowed us to improve some results. This research was supported by Program for the New Century Excellent Talents in University (NCET-09-0248), The PhD Programs Foundation of Ministry of Education of China (20100043110002), Fund of Jilin Provincial Science & Technology Department (No. 20111804) and the National Science Foundation of China.

Received 15 December 2011

REFERENCES

[1] Amato, U., Antoniadis, A. and Feis, I. D. (2006). Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis* **50** 2422–2446. MR2225577

[2] Aspirot, L., Bertin, K. and Perera, G. (2009). Asymptotic normality of the Nadaraya-Watson estimator for non-stationary functional data and applications to telecommunications. *Journal of Nonparametric Statistics* **21** 535–551. MR2543572

[3] Baíllo, A. and Grané, A. (2009). Local linear regression for functional predictor and scalar response. *Journal of Multivariate Analysis* **100** 102–111. MR2460480

[4] Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34** 2159–2179. MR2291496

[5] Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* **92** 24–41. MR2102242

[6] Efromovich, S. (1989). On sequential nonparametric estimation of a density. *Theory Probab. of Applications* **34**, 793–815. MR1005732

[7] Ferré, L. and Yao, A. F. (2003). Functional sliced inverse regression analysis. *Statistics* **37** 475–488. MR2022235

[8] Ferré, L. and Yao, A. F. (2005). Smoothed functional inverse regression. *Statistica Sinica* **15** 665–683. MR2233905

[9] Ferré, L. and Villa, N. (2006). Multilayer perceptron with functional inputs: An inverse regression approach. *Scandinavian Journal of Statistics* **33** 807–823. MR2300917

[10] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*, Springer, New York. MR2229687

[11] Ferraty, F., Manteiga, W. G., Calvo, A. M. and Vieu, P. (2011). Presmoothing in functional linear regression. *Statistica Sinica*. Preprint No: SS-10-085R2.

[12] Fung, W. K., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica* **12** 1093–1113. MR1947065

[13] Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projection from high dimensional data. *The Annals of Statistics* **21** 867–889. MR1232523

[14] Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics* **23** 73–102. MR1331657

[15] Huang, S. and Studden, W. Y. (1993). An equivalent kernel method for least squares spline regression. *Statist. Decisions* **3** 179–201. MR1244071
[16] James, G. and Sliverman, B. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association* **100** 565–576. MR2160560
[17] Manteiga, W. G. and Calvo, A. M. (2011). Bootstrap in functional linear regression. *Journal of Statistical Planning and Inference* **141** 453–461. MR2719509
[18] Müller, H. G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103** 426–437. MR2504202
[19] Leurgans, S., Moyeed, R. and Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society Series B* **55** 725–740. MR1223939
[20] Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics* **33** 1580–1616. MR2166556
[21] Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86** 316–342. MR1137117
[22] Ramsay, J. O. and Dalzeli, C. J. (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society Series B* **53** 539–572. MR1125714
[23] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*, 2nd ed. Springer, New York. MR1910407
[24] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, Springer, New York. MR2168993
[25] Wang, G., Lin, N. and Zhang, B. (2012). Functional linear regression after spline transformation. *Computational Statistics and Data Analysis* **56** 587–601. MR2853757
[26] Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100** 577–590. MR2160561
[27] Yao, F. and Müller, H. G. (2010). Functional quadratic regression. *Biometrika* **94** 49–64. MR2594416
[28] Yuan, M. and Cai, T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics* **38** 3412–3444. MR2766857
[29] Zhang, J. T. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics* **35** 1052–1079. MR2341698
[30] Zhou, J. (2006). Robust dimension reduction based on canonical correlation. *Journal of Multivariate Analysis* **100** 195–209. MR2460487
[31] Zhu, L. X., Miao, B. Q. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101** 630–643. MR2281245

Guochang Wang
Key Laboratory for Applied Statistics of MOE
and School of Mathematics and Statistics
Northeast Normal University
Changchun
China
E-mail address: wanggc023@nenu.edu.cn

Nan Lin
Department of Mathematics
Washington University in St. Louis
One Brookings Drive
Saint Louis, MO 63130
USA
E-mail address: nlin@math.wustl.edu

Baoxue Zhang
Key Laboratory for Applied Statistics of MOE
and School of Mathematics and Statistics
Northeast Normal University
Changchun
China
E-mail address: bxzhang@nenu.edu.cn