Functional Linear Regression: Dependence and Error Contamination

Cheng Chen[1], Shaojun Guo[2], and Xinghao Qiao[1]

[1]*Department of Statistics, London School of Economics, U.K.*

[2]*Institute of Statistics and Big Data, Renmin University of China, P.R. China*

Abstract

Functional linear regression is an important topic in functional data analysis. It is commonly assumed that samples of the functional predictor are independent realizations of an underlying stochastic process, and are observed over a grid of points contaminated by i.i.d. measurement errors. In practice, however, the dynamical dependence across different curves may exist and the parametric assumption on the error covariance structure could be unrealistic. In this paper, we consider functional linear regression with serially dependent observations of the functional predictor, when the contamination of the predictor by the white noise is genuinely functional with fully nonparametric covariance structure. Inspired by the fact that the autocovariance function of observed functional predictors automatically filters out the impact from the unobservable noise term, we propose a novel autocovariance-based generalized method-of-moments estimate of the slope function. We also develop a nonparametric smoothing approach to handle the scenario of partially observed functional predictors. The asymptotic properties of the resulting estimators under different scenarios are established. Finally, we demonstrate that our proposed method significantly outperforms possible competing methods through an extensive set of simulations and an analysis of a public financial dataset.

Some key words: Autocovariance; Eigenanalysis; Errors-in-predictors; Functional linear regression; Generalized method-of-moments; Local linear smoothing.

1 Introduction

In functional data analysis, the linear regression problem depicting the linear relationship between a functional predictor and either a scalar or functional response, has recently received a great deal of attention. See Ramsay and Silverman (2005) for a thorough discussion of the issues involved with fitting such data. For examples of recent research on functional linear models, see Yao et al. (2005); Hall and Horowitz (2007); Crambes et al. (2009); Cho et al. (2013); Chakraborty and Panaretos (2017) and the references therein. We refer to Morris (2015) for an extensive review on recent developments for functional regression.

In functional regression literature, one typical assumption is to model observed functional predictors, denoted by $X_1(\cdot), \ldots, X_n(\cdot)$, as independent realizations of an underlying stochastic process. However, curves can also arise from segments of consecutive measurements over time. Examples include daily curves of financial transaction data (Horvath et al., 2014), intraday electricity load curves (Cho et al., 2013) and daily pollution curves (Aue et al., 2015). Such type of curves, also named as curve time series, violates the independence assumption, in the sense that the dynamical dependence across different curves exists. The other key assumption treats the functional predictor as being either fully observed (Hall and Horowitz, 2007) or incompletely observed, with measurement error, at a grid of time points (Crambes et al., 2009). In the latter case, errors associated with distinct observation points are assumed to be i.i.d., where the corresponding covariance function for the error process is diagonal with constant diagonal components. In the curve time series setting, $X_t(\cdot)$ are often recorded at discrete points and are subject to dependent and heteroskedastic errors. Hence, the resulting error covariance matrix would be more nonparametric with varying diagonal entries and nonzero off-diagonal entries.

In this paper, we consider the functional linear regression in a time series context, which involves serially dependent observations of the functional predictor contaminated by gen-

uinely functional errors corresponding to a fully nonparametric covariance structure. We assume that the observed erroneous predictors, which we denote by $W_1(\cdot), \ldots, W_n(\cdot)$, are defined on a compact interval $\mathcal{U}$ and are subject to errors in the form of

$$W_t(u) = X_t(u) + e_t(u), \quad u \in \mathcal{U}, \tag{1}$$

where the error process $\{e_t(\cdot), t = 1, 2, \ldots\}$ is a sequence of white noise such that $E\{e_t(u)\} = 0$ for all t and $\mathrm{Cov}\{e_t(u), e_s(v)\} = 0$ for any $(u, v) \in \mathcal{U}^2$ provided $t \neq s$. We also assume that $X_t(\cdot)$ and $e_t(\cdot)$ are uncorrelated and correspond to unobservable signal and noise components, respectively. The error contamination model in (1) was also considered in Bathia et al. (2010). To fit the functional regression model, the conventional *least square* (LS) approach (Hall and Horowitz, 2007) relies on the sample covariance function of $W_t(\cdot)$, which is not a consistent estimator for the true covariance function of $X_t(\cdot)$, thus failing to account for the contamination that can result in substantial estimation bias. One can possibly implement the LS method in the resulting multiple linear regression after performing dimension reduction for $W_t(\cdot)$ to identify the dimensionality of $X_t(\cdot)$ (Bathia et al., 2010). However, this approach still suffers from unavoidable uncertainty due to $e_t(\cdot)$, while the inconsistency has been demonstrated by our simulations. Inspired from a simple fact that $\mathrm{Cov}\{W_t(u), W_{t+k}(v)\} = \mathrm{Cov}\{X_t(u), X_{t+k}(v)\}$ for any $k \neq 0$, which indicates that the impact from the unobservable noise term can be automatically eliminated, we develop an *autocovariance-based generalized method-of-moments* (AGMM) estimator for the slope function. This procedure makes the good use of the serial dependence information, which is the most relevant in the context of time series modelling.

To tackle the problem we consider, the conventional LS approach is not directly applicable in the sense that one cannot separate $X_t(\cdot)$ from $W_t(\cdot)$ in equation (1). This difficulty was resolved in Hall and Vial (2006) under the restrictive "low noise" setting, which assumes that the noise $e_t(\cdot)$ goes to zero as n grows to infinity. The recent work by Chakraborty and Panaretos (2017) implements the regression calibration approach combined with the low rank matrix completion technique to separate $X_t(\cdot)$ from $W_t(\cdot)$. Their approach relies on the identifiability result that, provided real analytic and banded covariance functions

for $X_t(\cdot)$ and $e_t(\cdot)$, respectively, the corresponding two covariance functions are identifiable (Descary and Panaretos, 2019). However, all the aforementioned methods are developed under the critical independence assumption, which would be inappropriate for the setting that $W_1(\cdot), \ldots, W_n(\cdot)$ are serially dependent.

The proposed AGMM method has four main advantages. First, it can handle regression with serially dependent observations of the functional predictor. The existence of dynamical dependence across different curves makes our problem tractable and facilitates the development of AGMM. Second, without placing any parametric assumption on the covariance structure of the error process, it relies on the autocovariance function to get rid of the effect from the genuinely functional error. Interestingly, it turns out that the operator in AGMM defined based on the autocovariance function of the curve process is identical to the non-negative operator in Bathia et al. (2010), which is used to assess the dimensionality of $X_t(\cdot)$ in equation (1). Third, the proposed method can be applied to both scalar and functional responses with either finite or infinite dimensional functional predictors. To handle a practical scenario where functional predictors are partially observed, we also develop a local linear smoothing approach. Theoretically we establish relevant convergence rates for our proposed estimators under different model settings. In particular, our asymptotic results for partially observed functional predictors reveal interesting phase transition phenomena. Fourth, empirically we illustrate the superiority of AGMM relative to the potential competitors.

The rest of the paper is organized as follows. In Section 2, we present the model for regression with dependent functional errors-in-predictors and develop AGMM fitting procedures for both scalar and functional responses. We also propose the regularized estimator by imposing some form of smoothness into the estimation procedure and discuss the selection of relevant tuning parameters. In Section 3, we present convergence results for our proposed estimators for the slope function under different functional scenarios. In Section 4, we develop a nonparametric smoothing approach for partially observed curve time series and investigate its asymptotic properties. Section 5 illustrates the finite sample performance of AGMM through a series of simulation studies and a public financial dataset. All technical

proofs are relegated to the Appendix and the Supplementary Material.

2 Methodology

2.1 Model setup

In this section, we describe the model setup for the functional linear regression with dependent errors-in-predictors we consider. Let $\mathcal{L}_2(\mathcal{U})$ denote a Hilbert space of square integrable functions defined on $\mathcal{U}$ equipped with the inner product $\langle f, g\rangle = \int_{\mathcal{U}} f(u)g(u)du$ for $f, g \in \mathcal{L}_2(\mathcal{U})$. Given a scalar response Y_t, a functional predictor $X_t(\cdot)$ in $\mathcal{L}_2(\mathcal{U})$, and, without loss of generality, assuming that $\{Y_t, X_t(\cdot)\}$ have been centered to have mean zero, the classical scalar-on-function linear regression model is of the form

$$Y_t = \int_{\mathcal{U}} X_t(u)\beta_0(u)du + \varepsilon_t, \quad t = 1, \ldots, n, \tag{2}$$

where the errors ε_t, independent of $X_{t+k}(\cdot)$ for any integer k, are generated according to a white noise process and $\beta_0(\cdot)$ is the unknown slope function. Generally, β_0 may not be uniquely determined. We will discuss how to identify β_0 we wish to estimate later.

We assume that the observed functional predictors $W_1(\cdot), \ldots, W_n(\cdot)$ satisfy the error contamination model in equation (1). The existence of the unobservable noise term $e_t(\cdot)$ indicates that the curves of interest, $X_t(\cdot)$, are not directly observed. Instead, they are recorded on a grid of points and are contaminated by the error process, $e_t(\cdot)$, without assuming any parametric structure on its covariance function, denoted by $C_e(u, v) = \text{Cov}\{e_t(u), e_t(v)\}$. This model guarantees that all the dynamic elements of $W_t(\cdot)$ are included in the signal term $X_t(\cdot)$ and all the white noise elements are absorbed into the noise term $e_t(\cdot)$. Furthermore, we assume that predictor errors $e_t(\cdot)$ are uncorrelated with both $X_{t+k}(\cdot)$ and ε_{t+k}, for all integer k.

Here we turn to discuss the identification of β_0. Assume that $\{(Y_t, X_t(\cdot))\}$ is strictly stationary and $C_0(u, v)$ is the covariance function of $X_t(\cdot)$, which admits the Karhunen-Loève expansion, $X_t(u) = \sum_{j=1}^{\infty} \xi_{tj}\phi_j(u)$, where $\xi_{tj} = \int_{\mathcal{U}} X_t(u)\phi_j(u)du$ and $\text{Cov}(\xi_{tj}, \xi_{tj'}) =$

$\lambda_j I(j = j')$ with $I(\cdot)$ denoting the indicator function. Then the eigenpairs $\{\lambda_j, \phi_j(\cdot)\}_{j\geqslant 1}$ satisfy the eigen-decomposition $\int_{\mathcal{U}} C_0(u,v)\phi_j(v)dv = \lambda_j\phi_j(u)$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$. Define $S_0(u) = E\{Y_t X_t(u)\}$, $d = \sup_{i\geqslant 1}\{i : \lambda_i > 0\}$ and assume $\sum_{j=1}^{d} \lambda_j^{-2}\{\mathrm{Cov}(Y_1, \xi_{1j})\}^2 < \infty$. Obviously β_0 satisfies the following equation

$$S_0(u) = \int_{\mathcal{U}} C_0(u,v)\beta(v)dv, u \in \mathcal{U}. \tag{3}$$

If the span of eigenfunctions $\{\phi_1, \ldots, \phi_d\}$ is dense in the $\mathcal{L}_2$ space, it is clear that β_0 is the unique solution to (3) and hence can be uniquely identified. In a general scenario, β_0 can also be well defined. To make β_0 identifiable, we consider the following minimization problem

$$\begin{aligned} &\min_{\beta\in\mathcal{L}^2(\mathcal{U})} \int_{\mathcal{U}} \beta^2(u)du, \\ &s.t.\ S_0(u) = \int_{\mathcal{U}} C_0(u,v)\beta(v)dv,\ u \in \mathcal{U}. \end{aligned} \tag{4}$$

Noting that the solution to (4) exists and is unique, we define the true slope function β_0 to be this unique minimizer in a closed form of $\beta_0 = \sum_{j=1}^{d} \lambda_j^{-1}\mathrm{Cov}(Y_1, \xi_{1j})\phi_j$, which holds for both $d < \infty$ and $d = \infty$. See also Cardot et al. (2003) and He et al. (2010).

2.2 Main idea

In this section, we describe the main idea to facilitate the development of AGMM to estimate $\beta_0(\cdot)$ in (2). We choose $X_{t+k}(\cdot)$ for $k = 0, 1, \ldots,$ as functional instrumental variables, which are assumed to be uncorrelated with the error ε_t in (2). Let

$$g_k^X(\beta, u) = \mathrm{Cov}\{Y_t, X_{t+k}(u)\} - \int_{\mathcal{U}} \mathrm{Cov}\{X_t(v), X_{t+k}(u)\}\beta(v)dv. \tag{5}$$

The population moment conditions, $E\{\varepsilon_t X_{t+k}(u)\} = 0$ for any $u \in \mathcal{U}$, and equation (2) implies that

$$g_k^X(\beta_0, u) \equiv 0 \text{ for any } u \in \mathcal{U} \text{ and } k = 1, \ldots. \tag{6}$$

In particular, the conventional LS approach is based on (6) with $k = 0$. However, this approach is inappropriate when $X_t(\cdot)$ are replaced by the surrogates $W_t(\cdot)$ given the fact that $C_W(u,v) = \mathrm{Cov}\{W_t(u), W_t(v)\} = C_0(u,v) + C_e(u,v)$, and hence the sample version of

$C_W(u,v)$ is not a consistent estimator for $C_0(u,v)$. See Hall and Vial (2006) for the identifiability of $C_0(u,v)$ and $C_e(u,v)$ under the assumption that the observed curves $W_1(\cdot),\ldots,W_n(\cdot)$ are independent and $e_t(\cdot)$ decays to zero as n goes to infinity.

To separate $X_t(\cdot)$ from $W_t(\cdot)$ under the serial dependence scenario, we develop a different approach without requiring the "low noise" condition. For an integer $k \geqslant 1$, denote the lag-k autocovariance function of $X_t(\cdot)$, by $C_k(u,v) = \mathrm{Cov}\{X_t(u), X_{t+k}(v)\}$, which does not depend on t. Our method is based on the simple fact that

$$\mathrm{Cov}\{Y_t, W_{t+k}(u)\} = \mathrm{Cov}\{Y_t, X_{t+k}(u)\} \text{ and } \mathrm{Cov}\{W_t(u), W_{t+k}(v)\} = C_k(u,v) \text{ for any } k \neq 0.$$

Then after substituting $X_t(\cdot)$ by $W_t(\cdot)$ in (5), we can also represent

$$g_k(\beta,u) = \mathrm{Cov}\big\{Y_t, W_{t+k}(u)\big\} - \int_{\mathcal{U}} \mathrm{Cov}\{W_t(v), W_{t+k}(u)\}\beta(v)dv = g_k^X(\beta,u),$$

and the moment conditions in (6) become

$$g_k(\beta_0,u) \equiv 0 \text{ for any } u \in \mathcal{U} \text{ and } k = 1\ldots,L,$$

where L is some prescribed positive integer.

Under the over-identification setting, where the number of moment conditions exceeds the number of parameters, we borrow the idea of *generalized methods-of-moments* (GMM) based on minimizing the distance from $g_1(\beta,\cdot),\ldots,g_L(\beta,\cdot)$ to zero. This distance is defined by the quadratic form of

$$Q(\beta) = \sum_{k=1}^{L}\sum_{l=1}^{L}\int_{\mathcal{U}}\int_{\mathcal{U}} g_k(\beta,u)\Omega_{k,l}(u,v)g_l(\beta,v)dudv,$$

where $\mathbf{\Omega}(u,v) = \{\Omega_{k,l}(u,v)\}_{1\leqslant k,l\leqslant L}$ is an L by L weight matrix whose (k,l)-th element is $\Omega_{k,l}(u,v)$. A suitable choice of $\mathbf{\Omega}(u,v)$ must satisfy the properties of symmetry and positive-definiteness (Guhaniyogi et al., 2013), which are, to be specific, (i) $\Omega_{kl}(u,v) = \Omega_{lk}(v,u)$ for each $k,l = 1,\ldots,L$ and $(u,v) \in \mathcal{U}^2$; (ii) for any finite collection of time points $u_1,\ldots,u_T$, $\sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbf{a}(u_t)^T\mathbf{\Omega}(u_t,u_{t'})\mathbf{a}(u_{t'})$ must be positive for any $\mathbf{a}(\cdot) = \big(a_1(\cdot),\ldots,a_L(\cdot)\big)^T$. In general, one can choose the optimal weight matrix $\mathbf{\Omega}$ and implement a two-step GMM. However, this would give a very slight improvement in our simulations. To simplify our derivation and

accelerate the computation, we choose the identity weight matrix as $\Omega_{k,l}(u,v) = I(k = l)I(u = v)$ and then minimize the resulting distance of

$$Q(\beta) = \sum_{k=1}^{L} \int_{\mathcal{U}} g_k(\beta, u)^2 du,$$

over $\beta(\cdot) \in \mathcal{L}_2(\mathcal{U})$. The minimizer of $Q(\beta)$, $\beta_0(\cdot)$, can be achieved by solving $\partial Q(\beta)/\partial\beta = 0$, i.e. for any $u \in \mathcal{U}$,

$$\sum_{k=1}^{L} \left[\int_{\mathcal{U}} C_k(u,z) \mathrm{Cov}\{Y_t, W_{t+k}(z)\} dz - \int_{\mathcal{U}} \Big\{ \int_{\mathcal{U}} C_k(u,z) C_k(v,z) dz \Big\} \beta(v) dv \right] = 0. \tag{7}$$

To ease our presentation, we define

$$R(u) = \sum_{k=1}^{L} \int_{\mathcal{U}} C_k(u,z) \mathrm{Cov}\{Y_t, W_{t+k}(z)\} dz \tag{8}$$

and

$$K(u,v) = \sum_{k=1}^{L} \int_{\mathcal{U}} C_k(u,z) C_k(v,z) dz. \tag{9}$$

Note that K can be viewed as the kernel of a linear operator acting on $\mathcal{L}_2(\mathcal{U})$, i.e. for any $f \in \mathcal{L}_2(\mathcal{U})$, K maps $f(u)$ to $\widetilde{f}(u) \equiv \int_{\mathcal{U}} K(u,v) f(v) dv$. For notational economy, we will use K to denote both the kernel and the operator. Indeed, the nonnegative definite operator K was proposed in Bathia et al. (2010) to identify the dimensionality of $X_t(\cdot)$ based on $W_t(\cdot)$ in (1). Substituting the relevant terms in (7), $\beta_0(\cdot)$ satisfies the following equation

$$R(u) = \int_{\mathcal{U}} K(u,v) \beta(v) dv \text{ for any } u \in \mathcal{U}. \tag{10}$$

See also functional extension of the least squares type of normal equation in (3).

Provided that $X_t(\cdot)$ is d-dimensional, it follows from Proposition 1 of Bathia et al. (2010) that, under regularity conditions, K has the spectral decomposition, $K(u,v) = \sum_{j=1}^{d} \theta_j \psi_j(u) \psi_j(v)$, with d nonzero eigenvalues $\theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_d$ and $\overline{\mathrm{span}}\{\psi_1, \ldots, \psi_d\}$ is the linear space spanned by the d eigenfunctions $\{\phi_1, \ldots, \phi_d\}$. This assertion still holds even for $d = \infty$.

Denote the null space of K and its orthogonal complement by $\ker(K) = \{x \in \mathcal{L}_2(\mathcal{U}) : Kx = 0\}$ and $\ker(K)^{\perp} = \{x \in \mathcal{L}_2(\mathcal{U}) : \langle x, y \rangle = 0, \forall y \in \ker(K)\}$, respectively. The inverse

operator K^{-1} corresponds to the inverse of the restricted operator $\breve{K} = K\big|\ker(K)^{\perp}$, which restricts the domain of K to $\ker(K)^{\perp}$. See Section 3.5 of Hsing and Eubank (2015) for details. When $d < \infty$, $\beta_0(\cdot)$ is indeed the unique solution to (10) in $\ker(K)^{\perp}$ in the form of

$$\beta_0(u) = \int_{\mathcal{U}} K^{-1}(u,v)R(v)dv = \sum_{j=1}^{d} \theta_j^{-1}\langle \psi_j, R\rangle \psi_j(u). \tag{11}$$

Provided K is a bounded operator when $d = \infty$, K^{-1} becomes an unbounded operator, which means it is discontinuous and cannot be estimated in a meaningful way. However, K^{-1} is usually associated with another function/operator, the composite function/operator can be reasonably assumed to be bounded, e.g. the regression operator (Li, 2018). If we further assume that the composite function $\int_{\mathcal{U}} K^{-1}(u,v)R(v)dv$ is bounded, or equivalently $\sum_{j=1}^{\infty} \theta_j^{-2}\langle \psi_j, R\rangle^2 < \infty$, $\beta_0(\cdot)$ is still the unique solution to (10) in $\ker(K)^{\perp}$ and is of the form

$$\beta_0(u) = \int_{\mathcal{U}} K^{-1}(u,v)R(v)dv = \sum_{j=1}^{\infty} \theta_j^{-1}\langle \psi_j, R\rangle \psi_j(u). \tag{12}$$

Both (11) and (12) motivate us to develop the estimation procedure for β_0 in Section 2.3.

2.3 Estimation procedure

In this section, we present the AGMM estimator for $\beta_0(\cdot)$ based on the main idea described in Section 2.2.

We first provide the estimates of $C_k(u,v)$ and $\mathrm{Cov}\{Y_t, W_{t+k}(u)\}$ for $k = 1, \ldots, L$, i.e.

$$\widehat{C}_k(u,v) = \frac{1}{n-L}\sum_{t=1}^{n-L} W_t(u)W_{t+k}(v) \quad \text{and} \quad \widehat{\mathrm{Cov}}\{Y_t, W_{t+k}(u)\} = \frac{1}{n-L}\sum_{t=1}^{n-L} Y_t W_{t+k}(u). \tag{13}$$

Combing (8), (9) and (13) gives the the natural estimators for $K(u,v)$ and $R(u)$ as

$$\widehat{K}(u,v) = \sum_{k=1}^{L}\int_{\mathcal{U}} \widehat{C}_k(u,z)\widehat{C}_k(v,z)dz = \frac{1}{(n-L)^2}\sum_{k=1}^{L}\sum_{t=1}^{n-L}\sum_{s=1}^{n-L} W_t(u)W_s(v)\langle W_{t+k}, W_{s+k}\rangle \tag{14}$$

and

$$\widehat{R}(u) = \sum_{k=1}^{L}\int_{\mathcal{U}} \widehat{C}_k(u,z)\widehat{\mathrm{Cov}}\{Y_t, W_{t+k}(z)\}dz = \frac{1}{(n-L)^2}\sum_{k=1}^{L}\sum_{t=1}^{n-L}\sum_{s=1}^{n-L} W_t(u)Y_s\langle W_{t+k}, W_{s+k}\rangle, \tag{15}$$

respectively. Note we choose a fixed integer $L > 1$, as K pulls together the information at different lags, while $L = 1$ may lead to spurious estimation results. See Section 2.5 for the discussion on the selection of L.

We next perform an eigenanalysis on $\widehat{K}$ and thus obtain the estimated eigenpairs $\{\widehat{\theta}_j, \widehat{\psi}_j(\cdot)\}$ for $j = 1, 2, \ldots$. When the number of functional observations n is large, the accumulated errors in (14), (15) and the eigenanalysis on $\widehat{K}$ are relatively small, thus resulting in smooth estimates of $\psi_j(\cdot)$ and $\beta_0(\cdot)$. We refer to this implementation of our method as Base AGMM for the remainder of the paper. However, in the setting without a sufficiently large n this version of AGMM suffers from a potential under-smoothing problem that the resulting estimate of $\beta_0(\cdot)$ wiggles quite a bit. To overcome this disadvantage, we can impose some level of smoothing in the eigenanalysis through the basis expansion approach, which converts the continuous functional eigenanalysis problem for $\widehat{K}$ to an approximately equivalent matrix eigenanalysis task. We explore this *basis expansion based AGMM*, simply referred to as AGMM from here on. To be specific, let $\mathbf{B}(u)$ be the J-dimensional orthnormal basis function, i.e. $\int_{\mathcal{U}} \mathbf{B}(u)\mathbf{B}^T(u)du = \mathbf{I}_J$, such that for each $j = 1, \ldots, J$, $\psi_j(\cdot)$ can be well approximated by $\boldsymbol{\delta}_j^T\mathbf{B}(\cdot)$, where $\boldsymbol{\delta}_j$ is the basis coefficients vector. Let

$$\widehat{\mathbf{K}} = \int_{\mathcal{U}}\int_{\mathcal{U}} \mathbf{B}(u)\mathbf{B}^T(v)\widehat{K}(u,v)dudv.$$

Performing an eigen-decomposision on $\widehat{\mathbf{K}}$ leads to the estimated eigenpairs $\{(\widehat{\theta}_j, \widehat{\boldsymbol{\delta}}_j)\}_{j=1}^J$. Then the j-th estimated principal component function is given by $\widehat{\psi}_j(\cdot) = \widehat{\boldsymbol{\delta}}_j^T\mathbf{B}(\cdot)$. See Section 2.5 for the selection of J. A similar basis expansion technique can be applied to produce a smooth estimate $\widehat{R}(\cdot)$. Note that $\widehat{\mathbf{K}}, \widehat{\theta}_j, \widehat{\psi}_j, j = 1, \ldots, d$, all depend on J, but for simplicity of notation, we will omit the corresponding superscripts where the context is clear.

Finally, we substitute the relevant terms in (11) and (12) by their estimated values. We discuss two situations corresponding to $d < \infty$ and $d = \infty$ as follows. (i) When $X_t(\cdot)$ is d-dimensional ($d < \infty$), we need to select the estimate $\widehat{d}$ of d in the sense that $\widehat{\theta}_1, \ldots, \widehat{\theta}_{\widehat{d}}$ are large eigenvalues of $\widehat{K}$ and $\widehat{\theta}_{\widehat{d}+1}$ drops dramatically. The estimate $\widehat{\beta}$ of β_0 is then given by

$$\widehat{\beta}(u) = \sum_{j=1}^{\widehat{d}} \widehat{\theta}_j^{-1}\langle\widehat{\psi}_j, \widehat{R}\rangle\widehat{\psi}_j(u). \tag{16}$$

(ii) When $X_t(\cdot)$ is an infinite dimensional functional object, we take the standard truncation approach by using the leading M eigenpairs of $\widehat{K}$ to approximate β_0 in (12). Specifically, we obtain the estimated slope function as

$$\widehat{\beta}(u) = \sum_{j=1}^{M} \widehat{\theta}_j^{-1} \langle \widehat{\psi}_j, \widehat{R} \rangle \widehat{\psi}_j(u). \tag{17}$$

Section 2.5 presents details to select $\widehat{d}$ and M. However, when $d = \infty$, the empirical performance of $\widehat{\beta}(\cdot)$ may be sensitive to the selected value of M. To improve the numerical stability, we suggest an alternative ridge-type method to estimate β_0. Specifically, we propose

$$\widehat{\beta}_{\text{ridge}}(u) = \sum_{j=1}^{\overline{M}} (\widehat{\theta}_j + \rho_n)^{-1} \langle \widehat{\psi}_j, \widehat{R} \rangle \widehat{\psi}_j(u), \tag{18}$$

where $\overline{M}$ is chosen to be reasonably larger than M and $\rho_n \geqslant 0$ is a ridge parameter. See also Hall and Horowitz (2007) for the ridge-type estimator in classical functional linear regression.

2.4 Generalization to functional response

In this section, we consider the case when the response is also functional. Given a functional response $Y_t(\cdot)$ and a functional predictor $X_t(\cdot)$, both of which are in $\mathcal{L}_2(\mathcal{U})$ and have mean zero, the function-on-function linear regression takes the form of

$$Y_t(u) = \int_{\mathcal{U}} X_t(v) \gamma_0(u, v) dv + \varepsilon_t(u), \ u \in \mathcal{U}, \ t = 1, \ldots, n, \tag{19}$$

where $\gamma_0(u, v)$ is the slope function of interest and $\varepsilon_t(\cdot)$, independent of $X_{t+k}(\cdot)$ for any integer k, are random elements in the underlying separable Hilbert space. We still observe the erroneous version $W_t(\cdot)$ rather than the signal $X_t(\cdot)$ itself in equation (1).

To estimate the slope function in (19), we develop an AGMM approach analogous to that for the scalar case in Section 2 by solving the normal equation of

$$H(u, v) = \int_{\mathcal{U}} K(u, w) \gamma(w, v) dw \text{ for any } v \in \mathcal{U}, \tag{20}$$

where $H(u, v) = \sum_{k=1}^{L} \int_{\mathcal{U}} C_k(u, z) \text{Cov}\{Y_t(v), W_{t+k}(z)\} dz$ with its natural estimator

$$\widehat{H}(u, v) = \frac{1}{(n-L)^2} \sum_{k=1}^{L} \sum_{t=1}^{n-L} \sum_{s=1}^{n-L} W_t(u) Y_s(v) \langle W_{t+k}, W_{s+k} \rangle. \tag{21}$$

Accordingly, we can provide the estimate $\widehat{\gamma}$ of γ_0 under two functional scenarios including $d < \infty$ and $d = \infty$. (i) When $d < \infty$, $\gamma_0(u, v)$ is the unique solution of (20) in $\ker(K)^{\perp}$ and can be represented as

$$\gamma_0(u, v) = \int_{\mathcal{U}} K^{-1}(u, w)H(w, v)dw = \sum_{j=1}^{d} \theta_j^{-1}\langle \psi_j, H(\cdot, v)\rangle \psi_j(u). \tag{22}$$

The estimate of $\gamma_0(u, v)$ is then given by

$$\widehat{\gamma}(u, v) = \sum_{j=1}^{\widehat{d}} \widehat{\theta}_j^{-1}\widehat{\psi}_j(u)\langle \widehat{\psi}_j, \widehat{H}(\cdot, v)\rangle. \tag{23}$$

(ii) Under the infinite dimensional setting ($d = \infty$, if we assume the boundedness of the composite function $\int_{\mathcal{U}} K^{-1}(u, w)H(w, v)dw$ in the L_2 sense, the solution to (20) uniquely exists. Approximating the infinite dimensional $\gamma_0(u, v)$ in (22) by the first M components and substituting the relevant terms by their estimated values, we can obtain

$$\widehat{\gamma}(u, v) = \sum_{j=1}^{M} \widehat{\theta}_j^{-1}\widehat{\psi}_j(u)\langle \widehat{\psi}_j, \widehat{H}(\cdot, v)\rangle. \tag{24}$$

2.5 Selection of tuning parameters

Implementing AGMM requires choosing L (selected lag length in (7)), M (truncated dimension in (17) when $d = \infty$), $\widehat{d}$ (number of identified nonzero eigenvalues of $\widehat{K}$ when $d < \infty$) and J (dimension of the basis function $\mathbf{B}(u)$). First, we tend to select a small value of L, as the strongest autocorrelations usually appear at the small time lags and adding more terms will make $\widehat{K}$ less accurate. Our simulated results suggest that the proposed estimators are not sensitive to the choice of L, therefore we set $L = 5$ in our empirical studies. See also Bathia et al. (2010) and Lam et al. (2011) for relevant discussions.

Second, to select M when $d = \infty$, the typical approach is to find the largest M eigenvalues of $\widehat{K}$ such that the corresponding cumulative percentage of variation exceeds the pre-specified threshold value, e.g. 90% or 95%. Other available methods include the bootstrap test (Bathia et al., 2010) and the eigen-ratio-based estimator (Lam et al., 2011). Third, to determine $\widehat{d}$ when $d < \infty$, we take the bootstrap approach proposed in Bathia et al. (2010). Our task

is to test the null hypothesis $H_0 : \theta_{d+1} = 0$. We reject H_0 if $\widehat{\theta}_{d+1} > c_\alpha$, where c_α is the critical value corresponding to the significant level $\alpha \in (0, 1)$. We summarize the bootstrap procedure as follows.

1. Define $\widehat{W}_t(\cdot) = \sum_{j=1}^{\widehat{d}} \widehat{\eta}_{tj}\widehat{\psi}_j(\cdot)$, where $\widehat{\eta}_{tj} = \int_{\mathcal{U}} W_t(u)\widehat{\psi}_j(u)du$ for $j = 1, \ldots, \widehat{d}$. Let $\widehat{e}_t(\cdot) = W_t(\cdot) - \widehat{W}_t(\cdot)$.

2. Generate a bootstrap sample using $W_t^*(\cdot) = \widehat{W}_t(\cdot) + e_t^*(\cdot)$, where e_t^* are drawn with replacement from $\{\widehat{e}_1, \ldots, \widehat{e}_n\}$.

3. In an analogy to $\widehat{K}$ defined in (14), form an estimator $\widehat{K}^*$ by replacing $\{W_t\}$ with $\{W_t^*\}$. Then calculate the $(d+1)$-th largest eigenvalue θ_{d+1}^* of $\widehat{K}^*$.

We repeat Steps 2 and 3 above B-times and reject H_0 if the event of $\{\widehat{\theta}_{d+1} > \theta_{d+1}^*\}$ occurs more than $[(1-\alpha)B]$ times. Starting with $\widehat{d} = 1$, we sequentially test $\theta_{\widehat{d}+1} = 0$ and increase $\widehat{d}$ by one until the resulting null hypothesis fails to be rejected.

Fourth, to select J, we propose the following G-fold cross-validation (CV) approach.

1. Sequentially divide the set $\{1, \ldots, n\}$ into G blockwise groups, $\mathcal{D}_1, \ldots, \mathcal{D}_G$, of approximately equal size.

2. Treat the g-th group as a validation set. Implement the regularized eigenanalysis in Section 2.3 on the remaining $G-1$ groups, compute $\widehat{\mathbf{K}}^{(-g)}$ and let $\widehat{\boldsymbol{\delta}}_1^{(-g)}, \ldots, \widehat{\boldsymbol{\delta}}_d^{(-g)}$ be the top d eigenvectors of $\widehat{\mathbf{K}}^{(-g)}$.

3. Compute $\widehat{K}^{(g)}(u, v)$ and $\widehat{\mathbf{K}}^{(g)}$ based on the validation set. Let $\widehat{\theta}_l^{(g)} = (\widehat{\boldsymbol{\delta}}_l^{(-g)})^T\widehat{\mathbf{K}}^{(g)}\widehat{\boldsymbol{\delta}}_l^{(-g)}$ for $l = 1, \ldots, d$.

We repeat Steps 2 and 3 above G times and choose J as the value that minimize the following mean CV error

$$\mathrm{CV}(J) = \frac{1}{G}\sum_{g=1}^{G}\int_{\mathcal{U}}\int_{\mathcal{U}}\Big\{\widehat{K}^{(g)}(u, v) - \sum_{j=1}^{d}\widehat{\theta}_j^{(g)}(\widehat{\boldsymbol{\delta}}_j^{(-g)})^T\mathbf{B}(u)\mathbf{B}(v)^T\widehat{\boldsymbol{\delta}}_j^{(-g)}\Big\}^2 dudv.$$

Given the time break on the training observations, the autocovariance assumption is jeopardized by $L = 5$ misutilized lagged terms. However, this effect on $\widehat{K}$ is negligible especially

when n is sufficiently large, hence our proposed CV approach can still be practically applied. See also Bergmeir et al. (2018) for various CV methods for time dependent data.

3 Theoretical properties

In this section, we investigate the theoretical properties of our proposed estimators for both scalar-on-function and function-on-function linear regressions.

To present the asymptotic results, we need the following regularity conditions.

Condition 1 *$\{W_t(\cdot), t = 1, 2, \dots\}$ is strictly stationary curve time series. Define the ψ-mixing with the mixing coefficients*

$$\psi(l) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{l}^{\infty}, P(A)P(B)>0} |1 - P(B|A)/P(B)|, \quad l = 1, 2, \dots,$$

where $\mathcal{F}_i^j$ denotes the σ-algebra generated by $\{W_t(\cdot), i \leqslant t \leqslant j\}$. Moreover, it holds that $\sum_{l=1}^{\infty} l\psi^{1/2}(l) < \infty$.

Condition 2 *$E(\|W_t\|^4) < \infty$ and $E(\varepsilon_t^2) < \infty$.*

The presentation of the ψ-mixing condition in Condition 1 is mainly for technical convenience. See Section 2.4 of Bosq (2000) on the mixing properties of curve time series. Condition 2 is the standard moment assumption in functional regression literature (Hall and Horowitz, 2007; Chakraborty and Panaretos, 2017).

Condition 3 *(i) When d is fixed, $\theta_1 > \cdots > \theta_d > 0 = \theta_{d+1}$; (ii) When $d = \infty$, $\theta_1 > \theta_2 > \cdots > 0$, and there exist some positive constants c and $\alpha > 1$ such that $\theta_j - \theta_{j+1} \geqslant cj^{-\alpha-1}$ for $j \geqslant 1$; (iii) $\overline{span}\{\phi_1, \dots, \phi_d\} = \overline{span}\{\psi_1, \dots, \psi_d\}$.*

Condition 4 *When $d = \infty$, $\beta_0(u) = \sum_{j=1}^{\infty} b_j\psi_j(u)$ and there exist some positive constants $\tau \geqslant \alpha + 1/2$ and C such that $|b_j| \leqslant Cj^{-\tau}$ for $j \geqslant 1$.*

Condition 3 restricts the eigen-structure of K and assumes that all the nonzero eigenvalues of K are distinct from each other. When $d = \infty$, Condition 3 (ii) prevents gaps

between adjacent eigenvalues from being too small. The parameter α determines the tightness of eigen-gaps with larger values of α yielding tighter gaps. This condition also indicates that $\theta_j \geqslant c\alpha^{-1}j^{-\alpha}$ as $\theta_j = \sum_{k=j}^{\infty}(\theta_k - \theta_{k+1}) \geqslant c\sum_{k=j}^{\infty} k^{-\alpha-1}$, and can be used to derive the convergence rates of estimated eigenfunctions. See also Hall and Horowitz (2007) and Qiao et al. (2019). Condition 4 restricts β_0 based on its expansion using eigenfunctions of K. The parameter τ determines the decay rate of slope basis coefficients, $\{b_j\}_{j=1}^{\infty}$. The assumption $\tau \geqslant \alpha + 1/2$ can be interpreted as requiring β_0 be sufficiently smooth relative to K, the smoothness of which can be implied by $\theta_j \geqslant c\alpha^{-1}j^{-\alpha}$ from Condition 3 (ii). See Hall and Horowitz (2007) for an analogous condition in functional linear regression.

Before presenting Theorem 1 for the asymptotic analysis of the scalar-on-function linear regression, we first solidify some notation. For any univariate function f, define $\|f\| = \sqrt{\langle f, f\rangle}$. We denote by $\|A\|_{\mathcal{S}}$ the Hilbert-Schmidt norm for any bivariate function A. The notation $a_n \asymp b_n$ for positive a_n and b_n means that the ratio a_n/b_n is bounded away from zero and infinity. To obtain $\widehat{\beta}$ in (16) when $d < \infty$, we use the consistent estimator for d defined as $\widehat{d} = \#\{j : \widehat{\theta}_j \geqslant \epsilon_n\}$, where ϵ_n satisfies the condition in Theorem 1 (i) below. Then by Theorem 3 of Bathia et al. (2010), $\widehat{d}$ converges in probability to d as $n \to \infty$.

Theorem 1 *Suppose that Conditions 1–4 hold. The following assertions hold as* $n \to \infty$:
(i) Let $\epsilon_n \to 0$ *and* $\epsilon_n^2 n \to \infty$ *as* $n \to \infty$. *When* d *is fixed, then*

$$\|\widehat{\beta} - \beta_0\| = O_P\big(n^{-1/2}\big).$$

(ii) When $d = \infty$, *if we further assume that* $M \asymp n^{1/(2\alpha+2\tau)}$, *then*

$$\|\widehat{\beta} - \beta_0\|^2 = O_P\big(M^{2\alpha+1}n^{-1} + M^{-2\tau+1}\big) = O_P\big(n^{-\frac{2\tau-1}{2\alpha+2\tau}}\big).$$

Remarks. (a) When d is fixed, the standard parametric root-n rate is achieved.

(b) When $d = \infty$, the convergence rate is governed by two sets of parameters (1) dimensionality parameter, sample size (n); (2) internal parameters, truncated dimension of the curve time series (M), decay rate of the lower bounds for eigenvalues (α), decay rate of the upper bounds for slope basis coefficients (τ). It is easy to see that larger values of α

(tighter eigen-gaps) yield a slower convergence rate, while increasing τ enhances the smoothness of $\beta_0(\cdot)$, thus resulting in a faster rate. The convergence rate consists of two terms, which reflects our familiar variance-bias tradeoff as commonly considered in nonparametric statistics. In particular, the bias is bounded by $O(M^{-\tau+1/2})$ and the variance is of the order $O_P(M^{2\alpha+1}n^{-1})$. To balance both terms, we choose the truncated dimension, $M \asymp n^{1/(2\alpha+2\tau)}$, while the optimal convergence rate then becomes $O_P\{n^{-(2\tau-1)/(2\alpha+2\tau)}\}$. It is also worth noting that this rate is slightly slower than the minimax rate $O_P\{n^{-(2\tau-1)/(\alpha+2\tau)}\}$ in Hall and Horowitz (2007), which considers independent observations of the functional predictor without any error contamination. In fact, we tackle a more difficult functional linear regression scenario, where extra complications come from the serial dependence and functional error contamination. From a theoretical perspective, whether the rate in part (ii) is optimal in the minimax sense is still of interest and requires further investigation.

Before presenting the asymptotic results for the function-on-function linear regression, we list Conditions 5 and 6 below, which are substitutes of Conditions 2 and 4, respectively, in the functional response case.

Condition 5 *$E(\|W_t\|^4) < \infty$ and $E(\|\varepsilon_t\|^2) < \infty$.*

Condition 6 *When $d = \infty$, $\gamma_0(u,v) = \sum_{j=1}^{\infty}\sum_{\ell=1}^{\infty} b_{j\ell}\psi_j(u)\psi_\ell(v)$ and there exist some positive constants $\tau \geqslant \alpha + 1/2$ and C such that $|b_{j\ell}| \leqslant C(j+\ell)^{-\tau-1/2}$ for $j, \ell \geqslant 1$.*

Theorem 2 *Suppose that Conditions 1, 3, 5 and 6 hold. The following assertions hold as $n \to \infty$:*

(i) Let $\epsilon_n \to 0$ and $\epsilon_n^2 n \to \infty$ as $n \to \infty$. When d is fixed, then

$$\|\widehat{\gamma} - \gamma_0\|_{\mathcal{S}} = O_P(n^{-1/2}).$$

(ii) When $d = \infty$, if we further assume that $M \asymp n^{1/(2\alpha+2\tau)}$, then

$$\|\widehat{\gamma} - \gamma_0\|_{\mathcal{S}}^2 = O_P\big(M^{2\alpha+1}n^{-1} + M^{-2\tau+1}\big) = O_P\big(n^{-\frac{2\tau-1}{2\alpha+2\tau}}\big).$$

4 Partially observed functional predictor

In this section, we consider a practical scenario where each $W_t(\cdot)$ is partially observed at random time points, $U_{t1}, \ldots, U_{tm_t} \in \mathcal{U} = [0,1]$, where for dense measurement designs all m_t's are larger than some order of n, and for sparse designs all m_t's are bounded (Zhang and Wang, 2016; Qiao et al., 2020). Let Z_{ti} represent the observed value of $W_t(U_{t_i})$ satisfying

$$Z_{ti} = W_t(U_{ti}) + \eta_{ti}, \quad i = 1, \ldots, m_t, \tag{25}$$

where η_{ti}'s are i.i.d. random errors with finite variance, independent of $W_t(\cdot)$.

Let $K(\cdot)$ be an univariate kernel function. We apply a local linear surface smoother to estimate the lag-k autocovariance function $C_k(u,v)$ for $k = 1, \ldots, L$ by minimizing

$$\sum_{t=1}^{n-L} \sum_{i=1}^{m_t} \sum_{j=1}^{m_{t+k}} \left\{ Z_{ti} Z_{(t+k)j} - a_0^{(k)} - a_1^{(k)}(U_{ti} - u) - a_2^{(k)}(U_{(t+k)j} - v) \right\}^2 K_{k,i,j,t,h}(u,v) \tag{26}$$

with respect to $(a_0^{(k)}, a_1^{(k)}, a_2^{(k)})$, where $K_{k,i,j,t,h}(u,v) = K\left(\frac{U_{ti}-u}{h_C}\right) K\left(\frac{U_{(t+k)j}-v}{h_C}\right)$ with a bandwidth $h_C > 0$. Let the minimizer of (26) be $(\widehat{a}_0^{(k)}, \widehat{a}_1^{(k)}, \widehat{a}_2^{(k)})$ and the resulting lag-k autocovariance estimator is $\widetilde{C}_k(u,v) = \widehat{a}_0^{(k)}$. Similarly, we implement a local linear smoothing approach to estimate $S_k(u) = \mathrm{Cov}(Y_t, W_{t+k}(u))$ for $k = 1, \ldots, L$ by minimizing

$$\sum_{t=1}^{n-L} \sum_{i=1}^{m_t} \left\{ Y_t Z_{(t+k)i} - b_0^{(k)} - b_1^{(k)}(U_{(t+k)i} - u) \right\}^2 K\left(\frac{U_{ti} - u}{h_S}\right) \tag{27}$$

with respect to $(b_0^{(k)}, b_1^{(k)})$ with a bandwidth $h_S > 0$. Then we obtain the estimate $\widetilde{S}_k(u) = \widehat{b}_0^{(k)}$. We also develop a basis expansion approach (Radchenko et al., 2015) to estimate C_k and S_k, where details can be found in Section C of the Supplementary Material.

Let $\widetilde{K}(u,v) = \sum_{k=1}^{L} \int_{\mathcal{U}} \widetilde{C}_k(u,z)\widetilde{C}_k(v,z)dz$ with estimated eigenpairs $(\widetilde{\theta}_j, \widetilde{\psi}_j)_{j \geqslant 1}$ and $\widetilde{R}(u) = \sum_{k=1}^{L} \int_{\mathcal{U}} \widetilde{C}_k(u,z)\widetilde{S}_k(z)dz$. In analogy to (16) and (17), we obtain the corresponding estimates $\widetilde{\beta}$ of β_0 by replacing $(\widehat{\theta}_j, \widehat{\psi}_j)_{j \geqslant 1}$ and $\widehat{R}$ with $(\widetilde{\theta}_j, \widetilde{\psi}_j)_{j \geqslant 1}$ and $\widetilde{R}$, respectively. Before presenting the main asymptotic results, we impose the following regularity conditions.

Condition 7 *(i) The errors $\{\eta_{ti}\}$ are i.i.d. mean zero random variables with $E|\eta_{ti}|^{2s} < \infty$ for some $s > 2$; (ii) $\{W_t(\cdot), t = 1, 2, \cdots\}$ is strictly stationary with ψ-mixing coefficients $\psi(l)$ satisfying $\psi(l) \lesssim l^{-\lambda}$ with $\lambda > \frac{3s-2}{s-2}$ and $\sup_{u \in [0,1]} E|W_t(u)|^{2s} < \infty$.*

Condition 8 *$K(\cdot)$ is a symmetric probability density function on $[-1,1]$ and is Lipschitz continuous.*

Condition 9 *$\{U_{ti}, i = 1,\ldots,m_t\}$ are i.i.d. copies of a random variable U defined on $[0,1]$ and the density $f(\cdot)$ of U is twice continuously differentiable and is bounded from below and above over $[0,1]$.*

Condition 10 *$\{W_t\}$ are independent of $\{U_{ti}\}$ and $\{\eta_{ti}\}$ are independent of $\{U_{ti}\}, \{W_t\}$.*

Condition 11 *(i) $\partial^2 C_k(u,v)/\partial u^2, \partial^2 C_k(u,v)/\partial u\partial v$ and $\partial^2 C_k(u,v)/\partial v^2$ for $k \geqslant 1$ are uniformly continuous and bounded on $[0,1]^2$; (ii) $\partial^2 S_k(u)/\partial u^2$ for $k \geqslant 1$ are uniformly continuous and bounded on $[0,1]$.*

Condition 12 *The number m_t of measurement locations in time t are independent random variables with distribution $m_t\rho_n^{-1} \sim \breve{m}$, where $\breve{m} \in \{1,\ldots,\bar{m}\}$ for some bounded $\bar{m}$ such that $P(\breve{m} > 1) > 0$.*

Condition 13 *The bandwidth parameters h_C and h_S satisfy*

$$h_C \to 0,\ h_S \to 0,\ \frac{\log(n\rho_n^2)}{(n\rho_n^2)^{\theta_C} h_C^2} \to 0 \text{ and } \frac{\log(n\rho_n)}{(n\rho_n)^{\theta_S} h_S} \to 0,$$

with

$$\theta_C = \frac{\beta - 2 - (1+\beta)/(s-1)}{\beta + 2 - (1+\beta)/(s-1)},\quad \theta_S = \frac{\beta - 3 - (1+\beta)/(s-1)}{\beta + 1 - (1+\beta)/(s-1)}.$$

Conditions 7–13 are standard in local linear smoothing when the serial dependence exists (Hansen, 2008; Rubín and Panaretos, 2020). In Condition 12, we treat the number m_t of measurement locations as random variables, but possibly diverges with n at the order of ρ_n. When ρ_n is bounded, it corresponds to the sparse case in Rubín and Panaretos (2020).

We present the convergence rates of $\widetilde{C}_k, \widetilde{S}_k$ for $k \geqslant 1$ and $\widetilde{\beta}_0$ in the following Theorems 3 and 4, respectively.

Theorem 3 *Suppose that Conditions 7–13 hold. As $n \to \infty$, we have*

$$\|\widetilde{C}_k - C_k\|_{\mathcal{S}} = O_P(\delta_{n1}) \text{ and } \|\widetilde{S}_k - S_k\| = O_P(\delta_{n2}) \text{ for } k \geqslant 1,$$

where

$$\delta_{n1} = \frac{1}{\sqrt{n\rho_n^2 h_C^2}} + \frac{1}{\sqrt{n}} + h_C^2 \text{ and } \delta_{n2} = \frac{1}{\sqrt{n\rho_n h_S}} + \frac{1}{\sqrt{n}} + h_S^2.$$

Theorem 4 *Suppose that Conditions 3–4 and 7–13 hold. The following assertions hold as* $n \to \infty$:

(i) Let $\epsilon_n \to 0$ *and* $\epsilon_n^2 n \to \infty$ *as* $n \to \infty$. *When* d *is fixed, then*

$$\|\widetilde{\beta} - \beta_0\| = O_P\big(\delta_{n1} + \delta_{n2}\big).$$

(ii) When $d = \infty$, *if we further assume that* $M \asymp \delta_{n1}^{-2/(2\alpha+2\tau)} + \delta_{n2}^{-2/(2\alpha+2\tau)}$, *then*

$$\|\widetilde{\beta} - \beta_0\|^2 = O_P\Big\{M^{2\alpha+1}\big(\delta_{n1}^2 + \delta_{n2}^2\big) + M^{-2\tau+1}\Big\} = O_P\Big\{\delta_{n1}^{\frac{2(2\tau-1)}{2\alpha+2\tau}} + \delta_{n2}^{\frac{2(2\tau-1)}{2\alpha+2\tau}}\Big\}.$$

Remarks. (a) In the sparse case where ρ_n is bounded, the L_2 rates of convergence for $\widetilde{C}_k$ and $\widetilde{S}_k$ in Theorem 3 become $O_P(n^{-1/2}h_C^{-1} + h_C^2)$ and $O_P(n^{-1/2}h_S^{-1/2} + h_S^2)$, respectively, which are consistent to those yielded convergence rates of one-dimensional and surface local linear smoothers for independent and sparsely sampled functional data (Zhang and Wang, 2016). When ρ_n grows with n, the convergence result reveals interesting phase transition phenomena depending on the relative order of ρ_n to n. We use different rates of $\widetilde{C}_k$ $(k \geqslant 1)$ to illustrate such phenomenon:

i. When $\rho_n/n^{1/4} \to 0$ with $n^{1/4}h \to \infty$, $\|\widetilde{C}_k - C_k\| = O_P(n^{-1/2}\rho_n^{-1}h_C^{-1} + h_C^2)$;

ii. When $\rho_n \asymp n^{1/4}$ with $h_C \asymp n^{-1/4}$ or $\rho_n/n^{1/4} \to \infty$ with $h_C = o(n^{-1/4})$ and $h_C\rho_n \to \infty$, $\|\widetilde{C}_k - C_k\| = O_P(n^{-1/2})$.

As ρ_n grows very fast, case (ii) results in the root-n rate, presenting that the theory for very dense curve time series falls in the parametric paradigm. As ρ_n grows moderately fast, case (i) corresponds to the rate faster than that for sparse data but slower than root-n. The rates under cases (i) and (ii) are respectively consistent to those of the estimated covariance function under categories of "dense" and and "ultra-dense" functional data (Zhang and Wang, 2016). For $\widetilde{S}_k$ $(k \geqslant 1)$, similar phase transition phenomenon occurs based on the ratio of ρ_n to $n^{1/4}$.

(b) The L_2 rates of $\widetilde{\beta}_0$ in Theorem 4 are governed by dimensionality parameters (n, ρ_n), bandwidth parameters (h_C, h_S) and those internal parameters in part (ii) of Theorem 1 when $d = \infty$. There also exists the phase transition based on the relative order of ρ_n to n. For example, when ρ_n is bounded and d is fixed, the rate of $\widetilde{\beta}_0$ is $O_P(n^{-1/2}h_C^{-1} + n^{-1/2}h_S^{-1/2} + h_C^2 + h_S^2)$. When ρ_n grows very fast with $\rho_n^{-1} = O(n^{-1/4})$ and suitable choices of h_C, h_S, the rates of $\widetilde{\beta}_0$ are identical to those for fully observed functional predictors in Theorem 1.

5 Empirical studies

5.1 Simulation study

In this section, we evaluate the finite sample performance of AGMM by a number of simulation studies. The observed predictor curves, $W_t(u), u \in [0, 1]$, are generated from equation (1) with

$$X_t(u) = \sum_{j=1}^{d} \xi_{tj}\phi_j(u) \quad \text{and} \quad e_t(u) = \sum_{j=1}^{10} \nu_{tj}\zeta_j(u),$$

where $\{\xi_{tj}\}_{t=1}^n$ follows a linear AR(1) process with the coefficient $(-1)^j(0.9 - 0.5j/d)$. The slope functions are generated by $\beta_0(u) = \sum_{j=1}^d b_j\phi_j(u)$, where b_j's take values from the first d components in $(2, 1.6, -1.2, 0.8, -1, -0.6)$. We generate responses $Y_1, \ldots, Y_n$ from equation (2), where ε_t are independent $N(0, 1)$ variables. Finally, we consider two different scenarios to generate $\{\phi_j(\cdot)\}_{j=1}^d$, $\{\zeta_j(\cdot)\}_{j=1}^{10}$ and $\{\nu_{tj}\}_{n\times 10}$.

Example 1: This example is taken from Bathia et al. (2010) with

$$\phi_j(u) = \sqrt{2}\cos(\pi j u), \quad \zeta_j(u) = \sqrt{2}\sin(\pi j u),$$

and the innovations ν_{tj} being independent standard normal variables.

We compare two versions of AGMM with three competing methods: covariance-based LS (CLS), covariance-based GMM (CGMM), autocovariance-based LS (ALS). The three competing approaches are implemented as follows. In the first two methods, we perform eigenanalysis on the estimated covariance function $\widehat{C}_W$, which converts the functional linear regression to the multiple linear regression, and then implement either LS or GMM. The

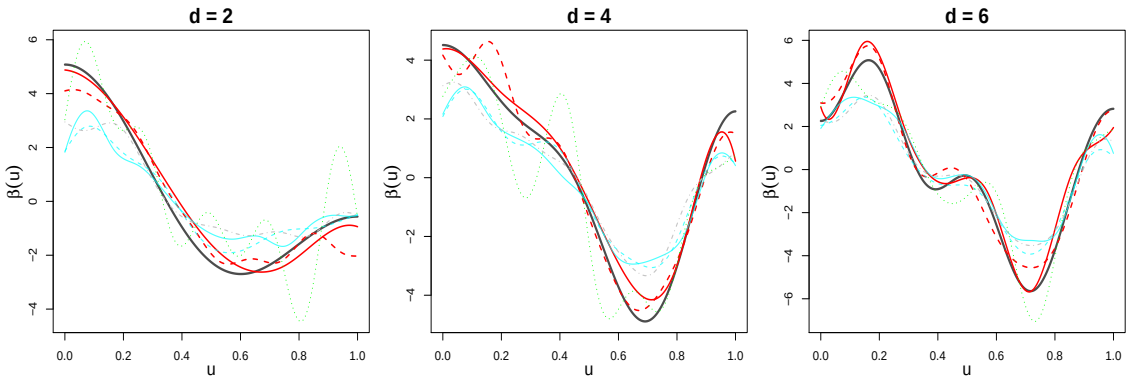

Figure 1: Example 1 with $n = 800$ and $d = 2, 4, 6$: Comparison of true $\beta(\cdot)$ functions (black solid) with median estimates over 100 simulation runs for AGMM (red solid), Base AGMM (red dashed), CLS (cyan solid), Base CLS (cyan dashed), Base CGMM (green dotted) and Base ALS (gray dash-dotted).

truncated dimension was chosen such that the selected principal components can explain more than 90% of the variation in the trajectory. We also tried the bootstrap method in Hall and Vial (2006) or to set a larger threshold level, e.g. 95%. However neither approach performed well, so we do not report the results here. The third ALS method relies on the eigenanalysis on the estimated autocovariance-based $\widehat{K}$ and the subsequent implementation of LS. In a similar fashion to the difference between Base AGMM and AGMM, we refer to each of the unregularized method as the "base" version.

The performance of four types of approaches are examined based on the mean integrated squared error for $\widehat{\beta}(u)$, i.e. $E[\int\{\widehat{\beta}(u) - \beta_0(u)\}^2 du]$. We consider different settings with $d = 2, 4, 6$ and $n = 200, 400, 800$, and ran each simulation 100 times. The regularized versions of CGMM and ALS did not give improvements in our simulation studies, so we do not report their results here. Figure 1 provides a graphical illustration of the results for $n = 800$ and $d = 2, 4, 6$. The black solid lines correspond to the true $\beta(u)$ from which the data were generated. The median most accurate estimate is also plotted for each of the competing methods. It is easy to see that the AGMM methods apparently provide the highest level of accuracy. The top part of Table 1 reports numerical summaries for all simulation scenarios. We can observe that the advantage of AGMM over Base AGMM is prominent especially

when either d or n is relatively small, while AGMM methods are superior to the competing methods when $n = 400$ or 800. However, under the setting with $n = 200$ and $d = 4$ or 6, the bootstrap test in Section 2.5 could not select $\widehat{d}$ very accurately, thus resulting in AGMM estimates inferior to some competitors.

Table 1: *Example 1*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs. The lowest values are in bold font.

$\widehat{d}$	n	d	Base CLS	CLS	Base CGMM	Base ALS	Base AGMM	AGMM
		2	1.320(0.026)	1.315(0.025)	2.215(0.099)	1.619(0.044)	1.187(0.052)	**0.720(0.033)**
	200	4	1.360(0.028)	**1.340(0.028)**	2.128(0.093)	2.451(0.102)	2.053(0.117)	1.704(0.107)
		6	1.337(0.030)	**1.320(0.029)**	1.912(0.102)	2.150(0.092)	1.847(0.098)	1.612(0.072)
		2	1.184(0.018)	1.181(0.019)	1.891(0.090)	1.338(0.026)	0.772(0.034)	**0.498(0.028)**
Est	400	4	1.198(0.021)	1.199(0.021)	1.939(0.090)	1.316(0.028)	0.701(0.034)	**0.584(0.034)**
		6	1.159(0.023)	1.154(0.022)	1.519(0.087)	1.323(0.034)	0.824(0.045)	**0.745(0.037)**
		2	1.159(0.012)	1.158(0.012)	1.792(0.080)	1.161(0.013)	0.346(0.013)	**0.211(0.012)**
	800	4	1.161(0.014)	1.160(0.014)	1.762(0.105)	1.122(0.014)	0.336(0.015)	**0.247(0.012)**
		6	1.123(0.014)	1.122(0.014)	1.297(0.091)	1.119(0.016)	**0.348(0.016)**	0.350(0.018)
		2	1.402(0.032)	1.238(0.030)	0.774(0.044)	1.637(0.044)	1.196(0.052)	**0.718(0.033)**
	200	4	1.365(0.030)	1.191(0.029)	0.924(0.056)	1.515(0.043)	1.214(0.071)	**0.797(0.046)**
		6	1.345(0.028)	1.272(0.027)	**1.150(0.065)**	1.465(0.036)	1.378(0.070)	1.196(0.057)
		2	1.226(0.019)	1.145(0.019)	0.503(0.027)	1.336(0.026)	0.772(0.034)	**0.498(0.028)**
True	400	4	1.199(0.021)	1.139(0.021)	0.529(0.024)	1.237(0.022)	0.653(0.032)	**0.488(0.029)**
		6	1.166(0.023)	1.139(0.022)	**0.656(0.038)**	1.170(0.023)	0.726(0.039)	0.704(0.042)
		2	1.174(0.012)	1.136(0.012)	0.269(0.011)	1.161(0.013)	0.346(0.013)	**0.211(0.012)**
	800	4	1.165(0.014)	1.131(0.014)	0.324(0.014)	1.130(0.014)	0.333(0.015)	0.**245(0.012)**
		6	1.121(0.014)	1.119(0.014)	**0.323(0.016)**	1.106(0.015)	0.336(0.015)	0.334(0.016)

To investigate the performance of AGMM after excluding the negative impact from the low accuracy of $\widehat{d}$ especially when $n = 200$, we also implement an "oracle" version, which uses the true d in the estimation. The numerical results are reported in the bottom part of Table 1. We can observe that GMM methods are superior to their LS versions, while CGMM slightly outperforms AGMM. These observations are due to the facts that, (i) top d eigenvalues for C_W and K correspond to the same signal components in Example 1, (ii)

GMM methods are capable of removing the impact from the noise term, (iii) the estimate $\widehat{C}_W$ in CGMM does not consider the functional error, while $\widehat{K}$ in AGMM would suffer from error accumulations. To better demonstrate the superiority of AGMM, we explore Example 2 below, where the covariance-based approach would fail to identify the signal components but its autocovariance-based version could.

Example 2: We generate $\{\zeta_j(\cdot)\}_{j=1}^{10}$ from a 10-dimensional orthonormal Fourier basis function, $\{\sqrt{2}\cos(2\pi ju), \sqrt{2}\sin(2\pi ju)\}_{j=1}^{5}$, and set $\phi_j(u) = \zeta_j(u)$ for $j = 1, \ldots, d$. The innovations ν_{tj} are independently sampled from $N(0, \sigma_j^2)$ with

$$\sigma_j^2 = \begin{cases} (1/2)^{j-1}, \text{ for } j = 1, \ldots, 6, \\ (2.6 - 0.1j) \times 1.1^{(d/2-3)}, \text{ for } j = 7, \ldots, 10. \end{cases}$$

In this example, provided the fact that $\{\phi_j(\cdot)\}_{j=1}^{d}$ shares the common basis functions with the first d elements in $\{\zeta_j(\cdot)\}_{j=1}^{10}$, we can calculate the variation in the trajectory explained by each of the 10 components under the population level. See Table 5 of the Supplementary Material for details. Take $d = 4$ as an illustrative example, the autocovariance-based methods can correctly identify the 4 signal components, while CLS and CGMM would mis-identify "7" and "8" as the signal components. Table 2 gives numerical summaries under the "oracle" scenario with true d in the estimation. As we would expect, two versions of AGMM provide substantially improved estimates, while Base AGMM is outperformed by AGMM in most of the cases. Under the scenario that $\widehat{d}$ is selected by the bootstrap approach, Figure 2 and Table 2 provide the graphical and numerical results, respectively. We observe similar trends as in Figure 1 and Table 1 with AGMM methods providing highly significant improvements over all the competitors.

Example 3: We use this example to demonstrate the sample performance of our proposed kernel smoothing approach to handle partially observed functional predictors. In each simulated scenario, we first generate $\{W_t(\cdot)\}$ and $\{e_t(\cdot)\}$ in the same way as Example 2 and then generate the observed values Z_{ti} from equation (25), where time points U_{ti} and errors η_{ti} are randomly sampled from Uniform$[0, 1]$ and $N(0, 0.5^2)$, respectively. We consider simulation settings $d = 2, 4, 6$, $n = 400, 800, 1200$ and $m_t = 10, 25, 50, 100$, changing from

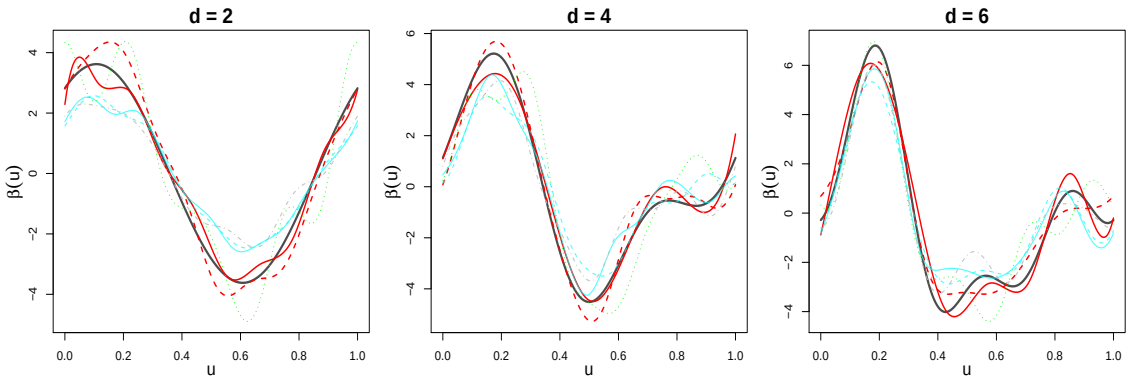

Figure 2: Example 2 with $n = 800$ and $d = 2, 4, 6$: Comparison of true $\beta(\cdot)$ functions (black solid) with median estimates over 100 simulation runs for AGMM (red solid), Base AGMM (red dashed), CLS (cyan solid), Base CLS (cyan dashed), Base CGMM (green dotted) and Base ALS (gray dash-dotted).

sparse to moderately dense to very dense measurement schedules. In each case, the optimal bandwidth parameters, h_C, h_S, are selected by the 10-fold cross-validation in Rubín and Panaretos (2020) and $\widehat{d}$ is chosen so that the first $\widehat{d}$ eigenvalues explains over 95% of the total variation. Table 3 reports numerical summaries for all 36 cases. Several conclusions can be drawn. First, for each d, the estimation accuracy is improved as n and m_t increase. Second, as curves are very densely observed, e.g. $m_t = 100$, our proposed smoothing approach enjoys similar performance with AGMM in Table 2, providing empirical evidence to support our remark for Theorem 4 about the same convergence rate between very densely observed and fully observed functional scenarios.

5.2 Real data analysis

In this section, we illustrate the proposed AGMM using a public financial dataset. The dataset was downloaded from `https://wrds-web.wharton.upenn.edu/wrds` and consists of one-minute resolution prices of Standard & Poor's 500 index and inclusive stocks from $n = 251$ trading days in year 2017. The trading time (9:30-16:00) is then converted to minutes, $u \in [0, 390]$. Let $P_t(u_j)$ $(t = 1, \ldots, n, j = 1, \ldots, 390)$ be the price of a financial asset

Table 2: *Example 2*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs. The lowest values are in bold font.

$\widehat{d}$	n	d	Base CLS	CLS	Base CGMM	Base ALS	Base AGMM	AGMM
		2	1.591(0.059)	0.990(0.046)	1.118(0.078)	1.165(0.030)	0.599(0.038)	**0.262(0.026)**
	400	4	2.026(0.066)	1.590(0.070)	2.310(0.112)	0.972(0.033)	0.686(0.041)	**0.448(0.034)**
		6	2.310(0.069)	1.932(0.077)	2.722(0.104)	0.938(0.035)	0.825(0.042)	**0.676(0.048)**
		2	1.377(0.051)	0.940(0.038)	0.884(0.085)	0.994(0.019)	0.337(0.020)	**0.138(0.010)**
True	800	4	1.934(0.051)	1.526(0.054)	2.268(0.105)	0.685(0.016)	0.318(0.016)	**0.208(0.013)**
		6	2.160(0.056)	1.872(0.055)	2.859(0.138)	0.575(0.015)	**0.339(0.017)**	0.364(0.020)
		2	1.294(0.053)	0.980(0.048)	0.750(0.081)	0.900(0.013)	0.203(0.011)	**0.080(0.005)**
	1200	4	1.959(0.053)	1.524(0.058)	2.426(0.121)	0.582(0.009)	0.167(0.008)	**0.124(0.006)**
		6	2.270(0.048)	2.002(0.050)	3.092(0.113)	0.494(0.011)	**0.217(0.010)**	0.248(0.010)
		2	0.817(0.012)	0.818(0.012)	0.980(0.059)	1.141(0.026)	0.575(0.030)	**0.248(0.018)**
	400	4	1.037(0.043)	0.725(0.036)	1.319(0.070)	1.097(0.038)	0.773(0.042)	**0.584(0.038)**
		6	0.913(0.041)	**0.811(0.038)**	1.305(0.068)	1.164(0.050)	0.999(0.051)	0.955(0.053)
		2	0.795(0.010)	0.795(0.010)	0.899(0.055)	0.989(0.019)	0.333(0.020)	**0.138(0.009)**
Est	800	4	1.093(0.033)	0.768(0.035)	1.471(0.065)	0.682(0.016)	0.319(0.016)	**0.212(0.013)**
		6	0.859(0.041)	0.809(0.039)	1.139(0.061)	0.571(0.016)	**0.335(0.017)**	0.369(0.020)
		2	0.779(0.007)	0.780(0.007)	0.747(0.044)	0.898(0.012)	0.205(0.012)	**0.079(0.005)**
	1200	4	1.055(0.026)	0.815(0.032)	1.344(0.052)	0.580(0.009)	0.166(0.008)	**0.130(0.007)**
		6	0.813(0.029)	0.808(0.029)	1.159(0.058)	0.492(0.011)	**0.216(0.011)**	0.243(0.009)

at the j-th minute after the opening time on the t-th trading day. Denote the *cumulative intraday return* (CIDR) trajectory, in percentage, by $r_t(u_j) = 100\big[\log\{P_t(u_j)\} - \log\{P_t(u_1)\}\big]$ (Horvath et al., 2014). Let $r_{m,t}(u)$ be the CIDR curves of the Standard & Poor's 500 index.

We extend the standard *capital asset pricing model* (CAPM) [Chapter 5 of Campbell et al. (1997)] to the functional domain by considering the functional linear regression with functional errors-in-predictors as follows

$$y_t = \alpha + \int x_t(u)\beta(u)du + \varepsilon_t, \quad r_{m,t}(u) = x_t(u) + e_t(u), \ t = 1, \ldots, n, \ u \in [0, 390], \tag{28}$$

where $x_t(\cdot)$ and $e_t(\cdot)$ represent the signal and error components in $r_{m,t}(\cdot)$, respectively, and y_t is the intraday return of a specific stock on the t-th trading day. Note that the slope

Table 3: *Example 3*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs.

n	d	$m_t = 10$	$m_t = 25$	$m_t = 50$	$m_t = 100$
	2	0.906(0.052)	0.374(0.019)	0.296(0.015)	0.227(0.011)
400	4	1.238(0.046)	0.637(0.027)	0.593(0.045)	0.395(0.020)
	6	1.168(0.051)	1.092(0.031)	0.906(0.028)	0.721(0.027)
	2	0.571(0.030)	0.194(0.009)	0.155(0.008)	0.142(0.007)
800	4	0.804(0.030)	0.375(0.015)	0.329(0.023)	0.231(0.010)
	6	1.130(0.039)	0.835(0.029)	0.481(0.019)	0.360(0.013)
	2	0.317(0.017)	0.145(0.007)	0.124(0.006)	0.107(0.005)
1200	4	0.632(0.025)	0.226(0.008)	0.214(0.013)	0.150(0.007)
	6	1.043(0.031)	0.505(0.016)	0.311(0.010)	0.269(0.009)

parameter in the classical CAPM explains how strongly an asset return depends on the market portfolio. Analogously, $\beta(\cdot)$ in functional CAPM in (28) can be understood as the functional sensitivity measure of an asset return to the market CIDR trajectory.

Figure 3 plots the estimated $\beta(\cdot)$ functions using both AGMM and CLS for three large-cap-sector stocks, Adobe (ADBE), Johnson & Johnson (JNJ) and PepsiCo (PEP). A few trends are apparent. First, the AGMM estimates place more positive weights as u increases. This result seems reasonable given the fact that the daily most recent market price would contain the most information about the stock's closing price. Second, the CLS estimates first dip in the mid-morning and then start to increase until the end of the trading day. In general, the shapes of the estimated $\beta(\cdot)$ functions by either AGMM or CLS are quite similar across the three stocks.

To formulate a prediction problem, we treat CIDR trajectories of the same stock as that in (28) up to current time $T < 390$ as $r_{y,t}(u), u \in [0, T]$, where, e.g., $T = 375$ corresponds to 15 minutes prior to the closing time of the trading day. Then we construct the same functional linear model as (28) by replacing $r_{m,t}(\cdot)$ with $r_{y,t}(\cdot)$. To judge which method produces superior predictions, we implement a rolling procedure to calculate the mean squared

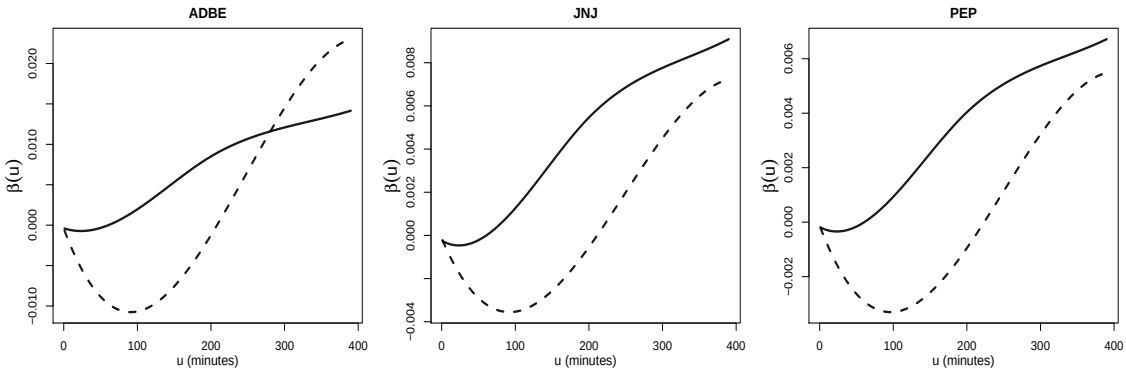

Figure 3: Estimated $\beta(\cdot)$ curves for AGMM (solid) and CLS (dashed).

prediction error (MSPE) for $H = 30$ days. Specifically, for each $h = H, H-1, \ldots, 1$, we treat $\{y_{n-h+1}, r_{y,n-h+1}\}$ as a testing set, implementing each fitting method on the training set of $\{(y_t, r_{y,t}) : t = 1, \ldots, n-h\}$, calculate the squared error between y_{n-h+1} and its predicted value, and repeat this procedure H-times to compute the MSPE. We calculate the MSPEs over a grid of (d, J) values and choose the pair with the lowest error. We also include the prediction errors from the null model, using the mean of the training response to predict the test response. The resulting MSPEs, for various values of T and the same three stocks, are provided in Table 4. It is easy to observe that the prediction accuracy for AGMM and CLS improves as T approaches to 390 and AGMM significantly outperforms two competitors in almost all settings.

Acknowledgements.

We are grateful to the editor, the associate editor and two referees for their insightful comments, which have led to significant improvement of our paper.

A Appendix

Appendices A.1 and A.2 contain proofs of Theorem 1 and Theorems 3–4. The proofs of Theorem 2 and all technical lemmas are in the Supplementary Material.

Table 4: Mean squared prediction errors up to different current times, $T = 330$, 345, 360, 375, 380 and 385 minutes, for AGMM and two competing methods. All entries have been multiplied by 10 for formatting reasons. The lowest MSPE for each value of T is bolded.

Stock	Method	$u \leqslant 330$	$u \leqslant 345$	$u \leqslant 360$	$u \leqslant 375$	$u \leqslant 380$	$u \leqslant 385$
	AGMM	1.276	**1.179**	**0.983**	**0.852**	**0.800**	**0.728**
ADBE	CLS	**1.272**	1.186	1.094	0.991	0.949	0.895
	Mean	12.224	12.224	12.224	12.224	12.224	12.224
	AGMM	**0.419**	**0.305**	**0.279**	**0.254**	**0.243**	**0.226**
JNJ	CLS	0.583	0.496	0.419	0.352	0.330	0.306
	Mean	3.077	3.077	3.077	3.077	3.077	3.077
	AGMM	**0.749**	**0.659**	**0.557**	**0.466**	**0.429**	**0.384**
PEP	CLS	0.781	0.687	0.596	0.502	0.468	0.429
	Mean	2.956	2.956	2.956	2.956	2.956	2.956

A.1 Proof of Theorem 1

A.1.1 Proof of Theorem 1 (i)

Define $\breve{K}(u,v) = \sum_{j=1}^{d} \widehat{\theta}_j \widehat{\psi}_j(u)\widehat{\psi}_j(v)$ and $K^{-1}(u,v) = \sum_{j=1}^{d} \theta_j^{-1}\psi_j(u)\psi_j(v)$. Let $\breve{\beta}(u) = \int_{\mathcal{U}} \breve{K}^{-1}(u,v)\widehat{R}(v)dv$. For a large $\delta > 0$, by Lemma 4, we have

$$\begin{aligned} P\big(n^{1/2}\|\widehat{\beta} - \beta_0\| > \delta\big) &= P\big(n^{1/2}\|\widehat{\beta} - \beta_0\| > \delta, \widehat{d} = d\big) + P\big(n^{1/2}\|\widehat{\beta} - \beta_0\| > \delta, \widehat{d} \neq d\big) \\ &\leqslant P\big(n^{1/2}\|\breve{\beta} - \beta_0\| > \delta, \widehat{d} = d\big) + P\big(\widehat{d} \neq d\big) \\ &\leqslant P\big(n^{1/2}\|\breve{\beta} - \beta_0\| > \delta\big) + o(1), \end{aligned}$$

which means that, to prove $n^{1/2}\|\widehat{\beta} - \beta_0\| = O_P(1)$, it suffices to show that $\|\breve{\beta} - \beta_0\| = O_P(n^{-1/2})$. It is easy to show that

$$\|\widetilde{\beta} - \beta_0\| \leqslant \|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}}\|\widehat{R}\| + \|K^{-1}\|_{\mathcal{S}}\|\widehat{R} - R\|. \tag{29}$$

Then it follows from Lemmas 2, 3 and 5 that $\|\breve{\beta} - \beta_0\| = O_P(n^{-1/2})$.

A.1.2 Proof of Theorem 1 (ii)

Without any ambiguity, write $\langle q, K\rangle$, $\langle K, q\rangle$ and $\langle p, \langle K, q\rangle\rangle$ for

$$\int_{\mathcal{U}} K(u,v)q(u)du, \int_{\mathcal{U}} K(u,v)q(v)dv \quad and \quad \int_{\mathcal{U}}\int_{\mathcal{U}} K(u,v)p(u)q(v)dudv,$$

respectively. In Lemma 6, we give expressions for $\widehat{\theta}_j - \theta_j$ and $\widehat{\psi}_j - \psi_j$ for $j \geqslant 1$.

Let $\beta_M(u) = \sum_{j=1}^{M} \theta_j^{-1}\langle \psi_j, R\rangle \psi_j(u)$. By the triangle inequality, we have

$$\|\widehat{\beta} - \beta_0\|^2 \leqslant \|\widehat{\beta} - \beta_M\|^2 + \|\beta_M - \beta_0\|^2. \tag{30}$$

By (12) and orthonormality of $\{\psi_j(\cdot)\}$, we have $\|\beta_M - \beta_0\|^2 = \sum_{j=M+1}^{\infty} \theta_j^{-2}\langle \psi_j, R\rangle^2$. It follows from Condition 4 and some specific calculations that

$$\|\beta_M - \beta_0\|^2 = \sum_{j=M+1}^{\infty} b_j^2 \leqslant C \sum_{j=M+1}^{\infty} j^{-2\tau} = O(M^{-2\tau+1}). \tag{31}$$

Next we will show the convergence rate of $\|\widehat{\beta} - \beta_M\|^2$. Observe that

$$\begin{aligned}
\widehat{\beta}(u) - \beta_M(u) &= \sum_{j=1}^{M} \big(\widehat{\theta}_j^{-1} - \theta_j^{-1}\big)\langle \psi_j, R\rangle \widehat{\psi}_j(u) + \sum_{j=1}^{M} \widehat{\theta}_j^{-1}\big(\langle \widehat{\psi}_j, \widehat{R}\rangle - \langle \psi_j, R\rangle\big)\widehat{\psi}_j(u) \\
&\quad + \sum_{j=1}^{M} \theta_j^{-1}\langle \psi_j, R\rangle\big\{\widehat{\psi}_j(u) - \psi_j(u)\big\}.
\end{aligned}$$

Then we have

$$\begin{aligned}
\|\widehat{\beta} - \beta_M\|^2 &\leqslant 3\sum_{j=1}^{M} \big(\widehat{\theta}_j^{-1} - \theta_j^{-1}\big)^2\langle \psi_j, R\rangle^2 + 3\sum_{j=1}^{M} \widehat{\theta}_j^{-2}\big(\langle \widehat{\psi}_j, \widehat{R}\rangle - \langle \psi_j, R\rangle\big)^2 \\
&\quad + 3M\sum_{j=1}^{M} \theta_j^{-2}\langle \psi_j, R\rangle^2 \big\|\widehat{\psi}_j - \psi_j\big\|^2 \\
&= 3I_{n1} + 3I_{n2} + 3I_{n3}.
\end{aligned} \tag{32}$$

Let $\widehat{\Delta} = \|\widehat{K} - K\|_{\mathcal{S}}$ and $\Omega_M = \{2\widehat{\Delta} \leqslant \delta_M\}$. On the event Ω_M, we can see that $\sup_{j \leqslant M} |\widehat{\theta}_j - \theta_j| \leqslant \theta_M/2$, which implies that $2^{-1}\theta_j \leqslant \widehat{\theta}_j \leqslant 2\theta_j$. Moreover, we can show that $P(\Omega_M) \to 1$ since $n^{1/2}\delta_M \to \infty$ as $n \to \infty$. Hence it suffices to work with bounds that are established under the event Ω_M.

Provided that event Ω_M holds, it follows from $\sup_{j\geqslant 1}|\widehat{\theta}_j-\theta_j| = O_P(n^{-1/2})$ in Lemma 1(i) and some calculations that

$$I_{n1} \leqslant 4\sum_{j=1}^{M}\big(\widehat{\theta}_j-\theta_j\big)^2\theta_j^{-4}\langle\psi_j,R\rangle^2 = 4\sum_{j=1}^{M}\theta_j^{-2}b_j^2\big(\widehat{\theta}_j-\theta_j\big)^2 = O_P\Big(n^{-1}\sum_{j=1}^{M}\theta_j^{-2}b_j^2\Big).$$

By Conditions 3–4, we have

$$I_{n1} = O_P(n^{-1})\cdot\Big(\sum_{j=1}^{M}j^{2\alpha-2\tau}\Big) = O_P(n^{-1})\cdot\big(M+M^{2\alpha-2\tau+1}\big) = o_P\big(n^{-1}M^{2\alpha+1}\big). \tag{33}$$

Consider the term I_{n3}. By $\|\widehat{\psi}_j-\psi_j\| = O_P\big(j^{1+\alpha}n^{-1/2}\big)$ in Lemma 1(iii) and Condition 4, we obtain that

$$I_{n3} \leqslant M\sum_{j=1}^{M}b_j^2\|\widehat{\psi}_j-\psi_j\|^2 = O_P\big(n^{-1}M^{2-2\tau+2\alpha+2}\big) = O_P\Big(n^{-1}M^{2\alpha+1}\Big), \tag{34}$$

where the last equality comes from $\alpha>1$ and $2\alpha-2\tau+4\leqslant 2\alpha+1$ implied by Condition 4.

Consider the term I_{n2}. On the event Ω_M, we have that

$$\begin{aligned}
I_{n2} &\leqslant 4\sum_{j=1}^{M}\theta_j^{-2}\big(\langle\widehat{\psi}_j,\widehat{R}\rangle-\langle\psi_j,R\rangle\big)^2\\
&\leqslant 12\sum_{j=1}^{M}\theta_j^{-2}\Big(\langle\widehat{\psi}_j-\psi_j,R\rangle^2+\langle\psi_j,\widehat{R}-R\rangle^2+\langle\widehat{\psi}_j-\psi_j,\widehat{R}-R\rangle^2\Big)\\
&\leqslant 12\sum_{j=1}^{M}\theta_j^{-2}\Big(\langle\widehat{\psi}_j-\psi_j,R\rangle^2+\|\widehat{R}-R\|^2+\|\widehat{\psi}_j-\psi_j\|^2\|\widehat{R}-R\|^2\Big),
\end{aligned} \tag{35}$$

where the last inequality comes from orthonormality of $\{\psi_j(\cdot)\}$ and Cauchy-Schwarz inequality. By Lemma 6 and some calculations, we can represent the term $\langle\widehat{\psi}_j-\psi_j,R\rangle$ as

$$\langle\widehat{\psi}_j-\psi_j,R\rangle = R_{j1}+R_{j2},$$

where $R_{j1}=\sum_{k:k\neq j}\theta_k b_k(\widehat{\theta}_j-\theta_k)^{-1}\langle\widehat{\psi}_j,\langle\widehat{K}-K,\psi_k\rangle\rangle$ and $R_{j2}=\theta_j b_j\langle\widehat{\psi}_j-\psi_j,\psi_j\rangle$. It follows from Condition 3–4, Lemma 1 and Cauchy-Schwarz inequality that

$$\sum_{j=1}^{M}\theta_j^{-2}R_{j2}^2 = O_P(n^{-1})\cdot\Big(\sum_{j=1}^{M}j^{-2\tau+2\alpha+2}\Big) = o_P\big(n^{-1}M^{2\alpha+1}\big). \tag{36}$$

Note that on the event Ω_M, $|\widehat{\theta}_j-\theta_j|\leqslant 2^{-1}|\theta_j-\theta_k|$ for $j=1,\ldots,k-1,k+1,\ldots,M$ and hence $|\widehat{\theta}_j-\theta_k|\geqslant 2^{-1}|\theta_j-\theta_k|$. If we can show that

$$\sup_{j\geqslant 1}(\theta_j^2 j^{2\alpha})^{-1}\sum_{k:k\neq j}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2} = O(1), \tag{37}$$

then, by Condition 4, Lemma 1 and on the event Ω_M, we have

$$\begin{aligned}\sum_{j=1}^{M}\theta_j^{-2}R_{j1}^2 &\leqslant 4\sum_{j=1}^{M}\theta_j^{-2}\sum_{k:k\neq j}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2}\|\widehat{K}-K\|_{\mathcal{S}}^2\\ &= O_P(n^{-1})\cdot\sum_{j=1}^{M}\theta_j^{-2}\theta_j^2 j^{2\alpha}=O_P(n^{-1}M^{2\alpha+1}).\end{aligned}\tag{38}$$

We turn to prove (37) as follows. Denote $[j/2]$ by the largest integer less than $j/2$. Then

$$\sum_{k:k\neq j}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2}=\left(\sum_{k=2(j+1)}^{\infty}+\sum_{k=[j/2]+1,k\neq j}^{k=2j+1}+\sum_{k=1}^{[j/2]}\right)\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2}.$$

Observe that for $k\geqslant 2(j+1)$,

$$\theta_j-\theta_k=\sum_{s=j}^{k-1}(\theta_s-\theta_{s+1})\geqslant c\int_{j+1}^{2(j+1)}s^{-\alpha-1}ds=-\frac{c}{\alpha}s^{-\alpha}\Big|_{j+1}^{2(j+1)}\geqslant\frac{c}{2\alpha}2^{-\alpha}j^{-\alpha},$$

and for $[j/2]+2\leqslant k\leqslant 2j+1$ but $k\neq j$,

$$|\theta_j-\theta_k|\geqslant\max(\theta_j-\theta_{j+1},\theta_{j-1}-\theta_j)\geqslant cj^{-\alpha-1}.$$

Therefore,

$$\begin{aligned}(\theta_j^2 j^{2\alpha})^{-1}\sum_{k=2(j+1)}^{\infty}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2} &= O(1)\cdot j^{2\alpha-2\tau}\sum_{k=2(j+1)}^{\infty}\theta_k^2=O(1),\\ (\theta_j^2 j^{2\alpha})^{-1}\sum_{k=[j/2]+1}^{2j+1}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2} &\leqslant (\theta_j^2 j^{2\alpha})^{-1}\sum_{k=[j/2]+1}^{2j+1}2\{\theta_j^2+(\theta_j-\theta_k)^2\}b_k^2(\theta_j-\theta_k)^{-2}\\ &= O(1)\cdot\theta_j^{-2}j^{-2\alpha}(1+\theta_j^2 j^{2\alpha+3-2\tau})=O(1),\\ (\theta_j^2 j^{2\alpha})^{-1}\sum_{k=1}^{[j/2]}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2} &\leqslant O(1)\sum_{k=1}^{[j/2]}\theta_k^2 b_k^2(\theta_k-\theta_{2k})^{-2}=O(1)\cdot\theta_1^2 j^{2\alpha-2\tau+1}=O(1),\end{aligned}$$

uniformly in j. Then (37) follows.

Moreover, it follows from Condition 3, Lemmas 1–3 that

$$\sum_{j=1}^{M}\theta_j^{-2}\|\widehat{R}-R\|^2=O_P(n^{-1}M^{2\alpha+1})\text{ and }\sum_{j=1}^{M}\theta_j^{-2}\|\widehat{\psi}_j-\psi_j\|^2\|\widehat{R}-R\|^2=O_P(n^{-2}M^{4\alpha+3}).\tag{39}$$

Combing the results in (35)–(36) and (38)–(39), we have

$$I_{n2}=O_P\Big(n^{-2}M^{4\alpha+3}+n^{-1}M^{2\alpha+1}\Big).\tag{40}$$

Combining the results in (30),(31) and (40) and choosing $M\asymp n^{1/(2\alpha+2\tau)}$, we obtain that

$$\|\widehat{\beta}-\beta_0\|^2=O_P\big(n^{-2}M^{4\alpha+3}+n^{-1}M^{2\alpha+1}+M^{-2\tau+1}\big)=O_P\big(n^{-\frac{2\tau-1}{2\alpha+2\tau}}\big).$$

A.2 Proofs of Theorems 3 and 4

Proof of Theorem 3. We begin with the L_2 rates of $\widetilde{C}_k$ for $k \geqslant 1$. We wish to prove them in the same fashion as the proof of Theorem 1 in Hansen (2008). For $p, q = 0, 1, 2$, define

$$\begin{aligned}
\widetilde{Z}^{(1)}_{p,q,i}(u,v) &= \sum_{i=1}^{m_t}\sum_{j=1}^{m_{t+k}} K_{k,i,j,h,t}(u,v)\left(\frac{U_{ti}-u}{h_C}\right)^p\left(\frac{U_{(t+k)j}-v}{h_C}\right)^q, \\
\widetilde{Z}^{(2)}_{p,q,i}(u,v) &= \sum_{i=1}^{m_t}\sum_{j=1}^{m_{t+k}} K_{k,i,j,h,t}(u,v)\left(\frac{U_{ti}-u}{h_C}\right)^p\left(\frac{U_{(t+k)j}-v}{h_C}\right)^q Z_{ti}Z_{(t+k)j}.
\end{aligned}$$

Let $S_{pq} = (n\rho_n^2 h_C^2)^{-1}\sum_{i=1}^n \widetilde{Z}^{(1)}_{p,q,i}$ and $G_{pq} = (n\rho_n^2 h_C^2)^{-1}\sum_{i=1}^n \widetilde{Z}^{(2)}_{p,q,i}$. Then we have

$$\widetilde{C}_k = \frac{(S_{20}S_{02}-S_{11}^2)G_{00} - (S_{10}S_{02}-S_{01}S_{11})G_{10} + (S_{10}S_{11}-S_{01}S_{20})G_{01}}{(S_{20}S_{02}-S_{11}^2)S_{00} - (S_{10}S_{02}-S_{01}S_{11})S_{10} + (S_{10}S_{11}-S_{01}S_{20})S_{01}}$$

so that $\widetilde{C}_k(u,v) - C_k(u,v)$ can be expressed as

$$\begin{aligned}
&= \frac{(S_{20}S_{02}-S_{11}^2)\{G_{00} - C_k(u,v)S_{00} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{10} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{01}\}}{(S_{20}S_{02}-S_{11}^2)S_{00} - (S_{10}S_{02}-S_{01}S_{11})S_{10} + (S_{10}S_{11}-S_{01}S_{20})S_{01}} \\
&- \frac{(S_{20}S_{02}-S_{11}^2)\{G_{10} - C_k(u,v)S_{10} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{20} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{11}\}}{(S_{20}S_{02}-S_{11}^2)S_{00} - (S_{10}S_{02}-S_{01}S_{11})S_{10} + (S_{10}S_{11}-S_{01}S_{20})S_{01}} \\
&+ \frac{(S_{20}S_{02}-S_{11}^2)\{G_{01} - C_k(u,v)S_{01} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{11} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{02}\}}{(S_{20}S_{02}-S_{11}^2)S_{00} - (S_{10}S_{02}-S_{01}S_{11})S_{10} + (S_{10}S_{11}-S_{01}S_{20})S_{01}}.
\end{aligned}$$

Let $\mathbb{U} = \{U_{ti}, i = 1, \ldots, m_t, t = 1, \ldots, n\}$. Suppose we have shown that for $p, q = 0, 1, 2$,

$$\left\|G_{pq} - E\{G_{pq}|\mathbb{U}\}\right\|_{\mathcal{S}} = O_P\left(\frac{1}{\sqrt{n\rho_n^2 h_C^2}} + \frac{1}{\sqrt{n}}\right), \tag{41}$$

and

$$\sup_{u,v\in[0,1]}\left|S_{pq}(u,v) - ES_{pq}(u,v)\right| = o_P(1). \tag{42}$$

By Taylor expansion, Condition 11 and (42),

$$\left\|E\{G_{00}|\mathbb{U}\} - C_k(u,v)S_{00} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{10} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{01}\right\|_{\mathcal{S}} = O_P(h_C^2). \tag{43}$$

Then combing (41) and (43) yields that

$$\left\|G_{00} - C_k(u,v)S_{00} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{10} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{01}\right\|_{\mathcal{S}} = O_P\left(\frac{1}{\sqrt{n\rho_n^2 h_C^2}} + \frac{1}{\sqrt{n}} + h_C^2\right). \tag{44}$$

Similarly, both $G_{10} - C_k(u,v)S_{10} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{20} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{11}$ and $G_{01} - C_k(u,v)S_{01} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{11} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{02}$ can be proved to have the same rate in (44). We can see from (42) that each denominator in $\widetilde{C}_k(u,v)$ is positive and bounded away from zero with probability approaching one, and as a consequence, part (i) of Theorem 3 follows.

Next, we turn to prove (41) and (42). For (41), it suffices to show that

$$\int\int E\left\{G_{00}(u,v) - E\big\{G_{00}(u,v)\big|\mathbb{U}\big\}\right\}^2 dudv \lesssim \frac{1}{n\rho_n^2 h_C^2} + \frac{1}{n},$$

where $a_n \lesssim b_n$ means $\limsup_{n\to\infty}|a_n/b_n| \leqslant C$ for some positive constant $C > 0$. It is easy to see that

$$\begin{aligned} E\left\{G_{00} - E\big\{G_{00}\big|\mathbb{U}\big\}\right\}^2 &\leqslant \frac{1}{n^2\rho_n^4 h_C^4}\sum_{t=1}^{n} E\left|\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}\big|\mathbb{U}\big\}\right|^2 \\ &+ \frac{1}{n\rho_n^4 h_C^4}\sum_{t=1}^{n}\left|\mathrm{Cov}\Big\{\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}\big|\mathbb{U}\big\}, \widetilde{Z}^{(2)}_{0,0,t+1} - E\big\{\widetilde{Z}^{(2)}_{0,0,t+1}\big|\mathbb{U}\big\}\Big\}\right|. \end{aligned}$$

Let $\mathbb{W} = \{W_t(\cdot), U_{ti}, i = 1,\ldots,m_t, t = 1,\ldots,n\}$. Note that $\widetilde{Z}^{(2)}_{0,0,1} - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\} = \widetilde{Z}^{(2)}_{0,0,1} - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{W}\} + E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{W}\} - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\}$. Given $\mathbb{W}$, the first term $\widetilde{Z}^{(1)}_{0,0,1} - E\{\widetilde{Z}^{(1)}_{0,0,1}|\mathbb{W}\}$ is a U-type statistics and hence some specific calculations yield that $E\Big\{\widetilde{Z}^{(2)}_{0,0,1} - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{W}\}\Big\}^2 \lesssim \rho_n^2 h_C^2 + \rho_n^3 h_C^3$. Moreover, $E\Big\{E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{W}\} - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\}\Big\}^2 \lesssim \rho_n^4 h_C^4 + \rho_n^2 h_C^2$. As a result,

$$E|\widetilde{Z}^{(2)}_{0,0,1} - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\}|^2 \lesssim \rho_n^4 h_C^4 + \rho_n^2 h_C^2.$$

In a similar manner together with Marcinkiewicz-Zygmund inequality, we can show that

$$E|\widetilde{Z}^{(2)}_{0,0,1}(u,v) - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\}|^s \lesssim \rho_n^{2s} h_C^{2s} + \rho_n^s h_C^s + \rho_n^{s/2} h_C^2.$$

For each fixed (u,v) and h_C, under Conditions 7–12, we see that $(\widetilde{Z}^{(2)}_{0,0,1}, \widetilde{Z}^{(2)}_{0,0,i}, \ldots)$ is strictly stationary with ψ-mixing coefficients $\psi_Z(l)$ satisfying $\psi_Z(l) \lesssim (l-k)^{-\lambda}$ for $l \geqslant k+1$. For $j^* < j \leqslant \max(j^*+1, \rho_n^{-2}h_C^{-2})$ with fixed $j^* > k+1$, we have that

$$\left|\mathrm{Cov}\big\{\widetilde{Z}^{(2)}_{0,0,1} - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\}, \widetilde{Z}^{(2)}_{0,0,t+1} - E\{\widetilde{Z}^{(2)}_{0,0,t+1}|\mathbb{U}\}\big\}\right| \lesssim \rho_n^4 h_C^4.$$

For $j > \max(j^*+1, \rho_n^{-2}h_C^{-2}) + 1$, using Davydov's lemma, we show that

$$\left|\mathrm{Cov}\big\{\widetilde{Z}^{(2)}_{0,0,1} - E\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\}, \widetilde{Z}^{(2)}_{0,0,t+1} - E\{\widetilde{Z}^{(2)}_{0,0,t+1}|\mathbb{U}\}\big\}\right| \lesssim j^{-2+2/s}\big(\rho_n^4 h_C^4 + \rho_n^2 h_C^2 + \rho_n h_C^{4/s}\big).$$

Therefore, the rate in (41) follows from the steps to prove Theorem 1 in Hansen (2008). Similarly, together with Conditions 7–13, the rates in (42) and $\|\widehat{S}_k - S_k\|$ follows from the steps to prove Theorem 2 in Hansen (2008). The proof is complete.

Proof of Theorem 4. By Theorem 3 for $k = 1, \ldots, L$, we can easily show that

$$\|\widetilde{K} - K\|_{\mathcal{S}} = O_P\big(\delta_{n1}\big) \text{ and } \|\widetilde{R} - R\| = O_P\big(\delta_{n1} + \delta_{n2}\big). \tag{45}$$

Following directly from the proof steps of Theorem 1 by replacing $\|\widehat{K} - K\|_{\mathcal{S}} = O_P(n^{-1/2})$ and $\|\widehat{R} - R\| = O_P(n^{-1/2})$ with the corresponding rates in (45), we complete our proof.

References

Aue, A., Norinho, D. and Hormann, S. (2015). On the prediction of stationary functional time series, *Journal of the American Statistical Association* **110**: 378–392.

Bathia, N., Yao, Q. and Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series, *The Annals of Statistics* **38**: 3352–3386.

Bergmeir, C., Hyndman, R. and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction, *Computational Statistics and Data Analysis* **120**: 70–83.

Bosq, D. (2000). *Linear Processes in Function Spaces - Theory and Applications*, Springer, New York.

Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*, Princeton University Press, New Jersey.

Cardot, H., Ferraty, F. and Sarda, P. (2003). Splines estimators for the functional linear model, *Statistica Sinica* **13**: 571–591.

Chakraborty, A. and Panaretos, V. M. (2017). Regression with genuinely functional errors-in-covariates, *arXiv:1712.04290.* .

Cho, H., Goude, Y., Brossat, X. and Yao, Q. (2013). Modeling and forecasting daily electricity load curves: a hybrid approach, *Journal of the American Statistical Association* **108**: 7–21.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression, *The Annals of Statistics* **37**: 35–72.

Descary, M.-H. and Panaretos, V. M. (2019). Functional data analysis by matrix completion, *Annals of Statistics* **47**: 1–38.

Guhaniyogi, R., Finley, A. O., Banerjee, S. and Kobe, R. (2013). Modeling complex spatial dependencies: low rank spatially varying cross-covariances with application to soil nutrient data, *Journal of Agricultural, Biological and Environmental Statistics* **18**: 274–298.

Hall, P. and Horowitz, J. Z. (2007). Methodology and convergence rates for functional linear regression, *The Annals of Statistics* **34**: 70–91.

Hall, P. and Vial, C. (2006). Assessing the finite dimensionality of functional data, *Journal of the Royal Statistical Society: Series B* **68**: 689–705.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data, *Econometric Theory* **24**: 726–748.

He, G., Mueller, H. G., Wang, J. L. and Yang, W. (2010). Functional linear regression via canonical analysis, *Bernoulli* **16**: 705729.

Horvath, L., Kokoszka, P. and Rice, G. (2014). Testing stationary of functional time series, *Journal of Econometrics* **179**: 66–82.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, John Wiley & Sons, Chichester.

Lam, C., Yao, Q. and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series, *Biometrika* **98**: 901–918.

Li, B. (2018). Linear operator-based statistical analysis: A useful paradigm for big data, *The Canadian Journal of Statistics* **46**: 79–103.

Morris, J. S. (2015). Functional regression, *Annual Review of Statistics and Its Application* **2**: 321–359.

Qiao, X., Guo, S. and James, G. (2019). Functional graphical models, *Journal of the American Statistical Association* **114**: 211–222.

Qiao, X., Qian, C., James, G. and Guo, S. (2020). Doubly functional graphical models in high dimensions, *Biometrika* **107**: 415431.

Radchenko, P., Qiao, X. and James, G. (2015). Index models for sparsely sampled functional data, *Journal of the American Statistical Association* **110**: 824–836.

Ramsay, J. and Silverman, B. (2005). *Functional data analysis (2nd ed.)*, Springer, New York.

Rubín, T. and Panaretos, V. M. (2020). Sparsely observed functional time series: Estimation and prediction, *Electronic Journal of Statistics* **14**: 1137–1210.

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

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond, *The Annals of Statistics* **44**: 2281–2321.

Supplementary Material to "Functional Linear Regression: Dependence and Error Contamination"

Cheng Chen, Shaojun Guo and Xinghao Qiao

This supplementary material contains proofs of Theorem 2 and all technical lemmas in Appendix B, the presentation of the basis expansion approach to address partially observed curve time series in Appendix C and additional simulation results in Appendix D.

B Additional technical proofs

B.1 Proof of Theorem 2

Following the similar arguments used in the proofs for Lemmas 2 and 3 under some regularity conditions, we can show that

$$\|\widehat{H} - H\|_{\mathcal{S}} = O_P(n^{-1/2}) \text{ and } \|H\|_{\mathcal{S}} = O(1). \tag{A.1}$$

Consider the case when d is fixed. Let $\widetilde{\gamma}(u,v) = \int_{\mathcal{U}} \check{K}^{-1}(u,w)\widehat{H}(w,v)dw$. Then we have

$$\|\widetilde{\gamma} - \gamma_0\|_{\mathcal{S}} \leqslant \|\check{K}^{-1} - K^{-1}\|_{\mathcal{S}}\|\widehat{H}\|_{\mathcal{S}} + \|K^{-1}\|_{\mathcal{S}}\|\widehat{H} - H\|_{\mathcal{S}}. \tag{A.2}$$

It follows from Lemma 5 and (A.1) that $\|\widetilde{\gamma} - \gamma\|_{\mathcal{S}} = O_P(n^{-1/2} + n^{-1/2}) = O_P(n^{-1/2})$. Finally, applying the similar technique used in the proof for part (i) of Theorem 1, we can prove the result in part (i) of Theorem 2.

When $d = \infty$, let $\gamma_M(u,v) = \sum_{j=1}^{M} \theta_j^{-1}\psi_j(u)\langle\psi_j, H(\cdot,v)\rangle$. By the triangle inequality, we have

$$\|\widehat{\gamma} - \gamma_0\|_{\mathcal{S}}^2 \leqslant \|\widehat{\gamma} - \gamma_M\|_{\mathcal{S}}^2 + \|\gamma_M - \gamma_0\|_{\mathcal{S}}^2. \tag{A.3}$$

It follows from Condition 6 and some specific calculations that

$$\begin{aligned} \|\gamma_M - \gamma_0\|_{\mathcal{S}}^2 &= O(1)\Big\| \sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty} b_{j\ell}\psi_j(u)\psi_\ell(v)\Big\|_{\mathcal{S}}^2 \\ &= O(1)\sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty} b_{j\ell}^2 = O(1)\sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty}(j+\ell)^{-2\tau-1} = O(M^{-2\tau+1}). \end{aligned} \tag{A.4}$$

It remains to show that the convergence rate of $\|\widehat{\gamma}-\gamma_M\|_{\mathcal{S}}^2$. Observe that

$$\begin{aligned}\widehat{\gamma}(u,v)-\gamma_M(u,v) &= \sum_{j=1}^{M}\big(\widehat{\theta}_j^{-1}-\theta_j^{-1}\big)\langle\psi_j,H\rangle(v)\widehat{\psi}_j(u)\\ &\quad+\sum_{j=1}^{M}\widehat{\theta}_j^{-1}\big(\langle\widehat{\psi}_j,\widehat{H}\rangle(v)-\langle\psi_j,H\rangle\big)(v)\widehat{\psi}_j(u)\\ &\quad+\sum_{j=1}^{M}\theta_j^{-1}\langle\psi_j,H\rangle(v)\big\{\widehat{\psi}_j(u)-\psi_j(u)\big\}.\end{aligned}$$

Then we have,

$$\begin{aligned}\|\widehat{\gamma}-\gamma_M\|_{\mathcal{S}}^2 &\leqslant 3\sum_{j=1}^{M}\big(\widehat{\theta}_j^{-1}-\theta_j^{-1}\big)^2\|\langle\psi_j,H\rangle\|^2+3\sum_{j=1}^{M}\widehat{\theta}_j^{-2}\|\langle\widehat{\psi}_j,\widehat{H}\rangle-\langle\psi_j,H\rangle\|^2\\ &\quad+3M\sum_{j=1}^{M}\theta_j^{-2}\|\langle\psi_j,H\rangle\|^2\|\widehat{\psi}_j-\psi_j\|^2.\end{aligned}$$

Following the similar arguments used in the proof for Theorem 1 (ii), we can show that

$$\|\widehat{\gamma}-\gamma_M\|_{\mathcal{S}}^2=O_P(M^{4\alpha+3}n^{-2}+M^{2\alpha+1}n^{-1}). \tag{A.5}$$

Combing the results in (A.3)–(A.5) and choosing $M\asymp n^{1/(2\alpha+2\tau)}$, we have

$$\|\widehat{\gamma}-\gamma_0\|_{\mathcal{S}}^2=O_P\big(M^{2\alpha+1}n^{-1}+M^{-2\tau+1}\big)=O_P\big(n^{-\frac{2\tau-1}{2\alpha+2\tau}}\big).$$

which completes our proof for part (ii) of Theorem 2.

B.2 Lemma 1 and its proof

Lemma 1 *Suppose that Conditions 1–3 hold and* $\langle\widehat{\psi}_j,\psi_j\rangle\geqslant 0$. *Then as* $n\to\infty$, *the following results hold:*

(i) $\|\widehat{K}-K\|_{\mathcal{S}}=O_P(n^{-1/2})$ *and* $\sup_{j\geqslant 1}|\widehat{\theta}_j-\theta_j|=O_P(n^{-1/2})$.

(ii) When d *is fixed,* $\|\widehat{\psi}_j-\psi_j\|=O_P(n^{-1/2})$ *for* $j=1,\ldots,d$.

(iii) When $d=\infty$, $\|\widehat{\psi}_j-\psi_j\|=O_P(j^{1+\alpha}n^{-1/2})$ *for* $j=1,2,\ldots$.

Proof. The first result in part (i) can be found in Theorem 1 of Bathia et al. (2010) and hence the proof is omitted. By (4.43) of Bosq (2000), we have $\sup_{j\geqslant 1}|\widehat{\theta}_j-\theta_j|\leqslant\|\widehat{K}-K\|_{\mathcal{S}}=$

$O_P(n^{-1/2})$, which completes the proof for the second result in part (i). To prove parts (ii) and (iii), let $\delta_j = 2\sqrt{2}\max\{(\theta_{j-1}-\theta_j)^{-1},(\theta_j-\theta_{j+1})^{-1}\}$ if $j \geqslant 2$ and $\delta_1 = 2\sqrt{2}(\theta_1-\theta_2)^{-1}$. It follows from Lemma 4.3 of Bosq (2000) that $\|\widehat{\psi}_j - \psi_j\| \leqslant \delta_j\|\widehat{K}-K\|_{\mathcal{S}} = O_P(\delta_j n^{-1/2})$. Under Condition 3(i) with a fixed d, root-n rate can be achieved. When $d=\infty$, Condition 3(ii) and (iii) imply that $\delta_j \leqslant Cj^{\alpha+1}$ with some positive constant C. This completes our proof for part (iii).

B.3 Lemma 2 and its proof

Lemma 2 *Suppose that Conditions 1-2 hold, then* $\|\widehat{R}-R\| = O_P(n^{-1/2})$.

Proof. Provided L is fixed, we may set $n \equiv n-L$. Let $\mathcal{S}$ denotes the space consisting of all the operators with a finite Hilbert-Schmidt norm and $\mathcal{H}$ denotes the space consisting of all the functions with a finite L_2 norm. Let $Z_{tk} = W_t \otimes W_{t+k} \in \mathcal{S}$ and $z_{tk} = Y_t W_{t+k} \in \mathcal{H}$. Now consider the kernel $\rho : \mathcal{S}\times\mathcal{H} \to \mathcal{H}$ given by $\rho(A,x) = Ax^*$ with $A \in \mathcal{S}$ and $x \in \mathcal{H}$. Let $c_k(\cdot) = \mathrm{Cov}\{Y_t, W_{t+k}(\cdot)\}$. We can represent $\widehat{C}_k\widehat{c}_k^* = n^{-2}\sum_{t=1}^n\sum_{t'=1}^n \rho(Z_{tk}, z_{t'k})$, which is simply a $\mathcal{H}$ valued Von Mises' functional (Borovskikh, 1996). For $d \geqslant 1$, neither of C_k and c_k is zero, it follows from Lemma 3 of Bathia et al. (2010) that $E\|\widehat{C}_k\widehat{c}_k^* - C_k c_k^*\|^2 = O(n^{-1})$. Then by the Chebyshev inequality, we have

$$\|\widehat{R}-R\| \leqslant \sum_{k=1}^{L}\|\widehat{C}_k\widehat{c}_k^* - C_k c_k^*\| = O_P(n^{-1/2}),$$

which completes the proof.

B.4 Lemma 3 and its proof

Lemma 3 *Suppose that Condition 2 holds, then* $\|R\| = O(1)$.

Proof. By the definitions of C_k and (8), we have

$$\|R\| \leqslant \sum_{k=1}^{L}\|C_k\|_{\mathcal{S}}\|\mathrm{Cov}(Y_t, W_{t+k})\| = \sum_{k=1}^{L}\|E\{W_t(u)W_{t+k}(v)\}\|_{\mathcal{S}}\|E(Y_t W_{t+k}(u))\|.$$

It follows from Cauchy-Schwartz inequality, Condition 2, Fubini Theorem and Jensen's inequality that $\|E\{W_t(u)W_{t+k}(v)\}\|_{\mathcal{S}}^2$

$$
\begin{aligned}
&= \int_{\mathcal{U}}\int_{\mathcal{U}}[E\{W_t(u)W_{t+k}(v)\}]^2 dudv \\
&\leqslant \int_{\mathcal{U}} E\{W_t(u)^2\}du \int_{\mathcal{U}} E\{W_{t+k}(v)^2\}dv = \Big[\int_{\mathcal{U}} E\{W_t(u)^2\}du\Big]^2 \leqslant E\Big\{\int_{\mathcal{U}} W_t(u)^2 du\Big\}^2 < \infty.
\end{aligned}
$$

Similarly, $\|E\{Y_t W_{t+k}(u)\}\|^2 \leqslant E(Y_t^2)\int_{\mathcal{U}} E\{W_{t+k}(u)^2\}du < \infty$. Combining the above results leads to $\|R\| = O(1)$.

B.5 Lemma 4 and its proof

Lemma 4 *Suppose the Conditions 1, 2, 3 (i) and (iii) hold. Let $\epsilon_n \to 0$, $\epsilon_n^2 n \to \infty$ and as $n \to \infty$. Then when $d < \infty$, $P\big(\widehat{d} \neq d\big) = O\{(\epsilon_n^2 n)^{-1}\} \to 0$.*

Proof. This lemma, which holds for $d < \infty$, can be found in Theorem 3 of Bathia et al. (2010) and hence the proof is omitted.

B.6 Lemma 5 and its proof

Lemma 5 *Suppose that Conditions 1, 2, 3(i) and (iii) hold. Then the following results hold.*
(i) $\|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}} = O_P(n^{-1/2})$.
(ii) $\|K^{-1}\|_{\mathcal{S}} = O(1)$.

Proof. Observe that

$$
\breve{K}^{-1} - K^{-1} = \sum_{j=1}^{d}(\widehat{\theta}_j^{-1} - \theta_j^{-1})\widehat{\psi}_j(u)\widehat{\psi}_j(v) + \sum_{j=1}^{d}\theta_j^{-1}\{\widehat{\psi}_j(u)\widehat{\psi}_j(v) - \psi_j(u)\psi_j(v)\}.
$$

Then by the orthonormality of $\{\psi_j(\cdot)\}$ and $\{\widehat{\psi}_j(\cdot)\}$, we have

$$
\|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}} \leqslant \sum_{j=1}^{d}\widehat{\theta}_j^{-1}\theta_j^{-1}|\widehat{\theta}_j - \theta_j| + 2\sum_{j=1}^{d}\theta_j^{-1}\|\widehat{\psi}_j - \psi_j\|. \tag{A.6}
$$

When d is fixed, the smallest eigenvalue θ_d is bounded away from zero. It follows from Lemma 1 (i),(ii) and (A.6) that there exists some positive constant C such that $\|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}} \leqslant C(\theta_d^{-2} + \theta_d^{-1})n^{-1/2}$, which completes the proof for part (i).

Note that $\|K^{-1}\|_{\mathcal{S}} = \|\sum_{j=1}^{d} \theta_j^{-1}\psi_j(u)\psi_j(v)\|_{\mathcal{S}} = (\sum_{j=1}^{d} \theta_j^{-2})^{1/2} \leqslant d^{1/2}\theta_d^{-1}$. Then part (ii) follows as d is fixed and θ_d is bounded below from zero.

B.7 Lemma 6 and its proof

Lemma 6 *If* $\inf_{k\neq j} |\widehat{\theta}_j - \theta_k| > 0$, *then*

$$\widehat{\psi}_j - \psi_j = \sum_{k:k\neq j} (\widehat{\theta}_j - \theta_k)^{-1}\psi_k\langle\hat{\psi}_j, \langle\widehat{K} - K, \psi_k\rangle\rangle + \psi_j\langle\widehat{\psi}_j - \psi_j, \psi_j\rangle. \tag{A.7}$$

Proof. This lemma can be derived from Lemma 5.1 of Hall and Horowitz (2007) and hence the proof is omitted.

C Basis expansion approach

We develop a standard basis expansion approach to estimate $K(u, v)$ and $R(u)$. Let $\mathbf{B}(u)$ be the J-dimensional orthonormal basis function, i.e. $\int_{\mathcal{U}} \mathbf{B}(u)\mathbf{B}^T(u)du = \mathbf{I}_J$, such that each $C_k(u, v)$ can be well approximated by $\{\mathbf{B}(u)\}^T\mathbf{\Sigma}_k\mathbf{B}(v)$. In practice, J can be selected by a similar cross-validation procedure described in Section 2.5. Let $\mathbf{B}_{ti} = \mathbf{B}(U_{ti})$. We consider minimizing

$$\sum_{t=1}^{n-L}\sum_{i=1}^{m_t}\sum_{j=1}^{m_{t+k}} \left\{Z_{ti}Z_{(t+k)j} - \mathbf{B}_{ti}^T\mathbf{\Sigma}_k\mathbf{B}_{(t+k)j}\right\}^2 \tag{A.8}$$

with respect to $\mathbf{\Sigma}_k \in \mathbb{R}^{J\times J}$. Standard calculation shows that the estimate of $\mathbf{\Sigma}_k$ that minimizes (A.8) is

$$\mathrm{vec}(\widehat{\mathbf{\Sigma}}_k) = \left(\sum_{t,i,j}(\mathbf{B}_{(t+k)j}\otimes\mathbf{B}_i)(\mathbf{B}_{(t+k)j}\otimes\mathbf{B}_i)^T\right)^{-1}\sum_{t,i,j}(\mathbf{B}_{(t+k)j}\otimes\mathbf{B}_i)Z_{ti}Z_{(t+k)j},$$

where $\mathrm{vec}(\mathbf{B})$ denotes the vectorization of the matrix $\mathbf{B}$ formed by stacking its columns into a single column vector and $\otimes$ is the Kronecker product. Then the estimate of $K(u, v)$ is

$$\widetilde{K}(u, v) = \{\mathbf{B}(u)\}^T\sum_{k=1}^{L}\widehat{\mathbf{\Sigma}}_k\widehat{\mathbf{\Sigma}}_k^T\mathbf{B}(v).$$

Similarly, we can obtain a consistent estimator $\widehat{\mathrm{Cov}}\{Y_t, W_{t+k}(u)\} = \widehat{\boldsymbol{\delta}}_k^T \mathbf{B}(u)$, where $\widehat{\boldsymbol{\delta}}_k$ is obtained by minimizing

$$\sum_{t=1}^{n-L} \sum_{1 \leqslant i \leqslant m_t} \left\{ Y_t Z_{(t+k)i} - \boldsymbol{\delta}_k^T \mathbf{B}_{(t+k)i} \right\}^2$$

with repsect to $\boldsymbol{\delta}_k \in \mathbb{R}^J$. Then the estimate of $\boldsymbol{\delta}_k$ is

$$\widehat{\boldsymbol{\delta}}_k = \left(\sum_{t,i} \mathbf{B}_{(t+k)i} \mathbf{B}_{(t+k)i}^T \right)^{-1} \sum_{t,i} \mathbf{B}_{(t+k)i} Y_t Z_{(t+k)i}.$$

As a result, $R(u)$ can be estimated by

$$\widetilde{R}(u) = \sum_{k=1}^{L} \widehat{\boldsymbol{\delta}}_k^T \widehat{\boldsymbol{\Sigma}}_k^T \mathbf{B}(u).$$

D Additional simulation results

For Example 2, Table 5 reports the variance explained by each of the 10 components under the population level. For each of the three parts corresponding to $d = 2, 4$ and 6, the second and third rows provide the variance explained by each of the d signal components and 10 error components, respectively. The first row ranks the components based on the overall variance explained by each individual component, where the fourth row displays the corresponding values. Take $d = 4$ as an illustrative example, the autocovariance-based approach can correctly identify the first four signal components, while the covariance-based approach can only correctly identify "1" and "2", but incorrectly select "7" and "8" as signal components. Moreover, we consider another scenario for Example 2 by generating innovations $\{\nu_{tj}\}$ from a standard normal distribution, where the variance decomposition is illustrated via Table 6. Under this setting, we can observe that both approaches are capable of correctly identifying the d signal components.

We next illustrate the sample performance of AGMM using two additional simulated examples to support Section 5.1.

Example 4: This example is used to demonstrate the superiority of AGMM methods under the setting where the dimension of the $\beta_0(\cdot)$ is less than the dimension of $X_t(\cdot)$. While

Table 5: The variance explained by each of the components in Example 2. Top d components identified by covaraicne-based and autocovariance-based approaches are underlined and in bold font, respectively.

	Component	1	2	7	8	9	10	3	4	5	6
	Signal	**1.73**	**1.19**								
d=2	Error	1.00	0.50	1.57	1.49	1.40	1.32	0.25	0.13	0.06	0.03
	Sum	2.73	1.69	1.57	1.49	1.40	1.32	0.25	0.13	0.06	0.03
	Component	1	2	7	8	9	10	3	4	5	6
	Signal	**2.50**	**1.73**					**1.38**	**1.19**		
d=4	Error	1.00	0.50	1.73	1.64	1.55	1.45	0.25	0.13	0.06	0.03
	Sum	3.50	2.23	1.73	1.64	1.55	1.45	1.63	1.32	0.06	0.03
	Component	1	2	3	7	8	9	10	4	5	6
	Signal	**3.00**	**2.16**	**1.73**					**1.47**	**1.30**	**1.19**
d=6	Error	1.00	0.50	0.25	1.90	1.80	1.70	1.60	0.13	0.06	0.03
	Sum	4.00	2.66	1.98	1.90	1.80	1.70	1.60	1.60	1.37	1.22

Table 6: The variance explained by each of the components in Example 2 with $\{\nu_{tj}\}$ being $N(0,1)$ variables. Top d components identified by covaraicne-based and autocovariance-based approaches are underlined and in bold font, respectively.

d=2	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**1.73**	**1.19**								
	Error	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Sum	2.73	2.19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d=4	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**2.50**	**1.73**	**1.38**	**1.19**						
	Error	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Sum	3.50	2.73	2.38	2.19	1.00	1.00	1.00	1.00	1.00	1.00
d=6	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**3.00**	**2.16**	**1.73**	**1.47**	**1.30**	**1.19**				
	Error	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Sum	4.00	3.16	2.73	2.47	2.30	2.19	1.00	1.00	1.00	1.00

the data are generated in the same fashion to Example 2, the slope functions are generated by $\beta_0(\cdot) = \sum_{j=1}^{d} \widetilde{b}_j \phi_j(\cdot)$ with $\widetilde{b}_j = b_j$ for $j = 1, \ldots, d-1$ and $\widetilde{b}_d = 0$ so that the dimension of $\beta_0(\cdot)$ is $d-1 < d$. Table 7 provides numerical results under the oracle scenario with true d in the estimation. We obtain the same findings to those in Table 2, i.e. two versions of AGMM significantly outperform their competing methods, while AGMM is superior to Base AGMM in most of the cases.

Table 7: *Example 4*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs. The lowest values are in bold font.

n	d	Base CLS	CLS	Base CGMM	Base ALS	Base AGMM	AGMM
	2	0.683(0.008)	0.577(0.007)	0.244(0.014)	0.646(0.008)	0.255(0.014)	**0.132(0.008)**
400	4	1.415(0.042)	0.993(0.043)	1.619(0.072)	0.756(0.014)	0.489(0.024)	**0.324(0.018)**
	6	1.990(0.051)	1.600(0.055)	2.312(0.066)	0.775(0.021)	0.647(0.025)	**0.500(0.024)**
	2	0.589(0.006)	0.560(0.006)	0.137(0.008)	0.593(0.006)	0.125(0.005)	**0.076(0.005)**
800	4	1.378(0.038)	0.855(0.037)	1.641(0.069)	0.620(0.008)	0.253(0.010)	**0.191(0.012)**
	6	1.817(0.036)	1.546(0.035)	2.351(0.077)	0.515(0.009)	**0.295(0.011)**	0.304(0.019)
	2	0.573(0.004)	0.552(0.004)	0.081(0.005)	0.576(0.004)	0.082(0.004)	**0.048(0.003)**
1200	4	1.383(0.035)	0.875(0.044)	1.732(0.072)	0.554(0.005)	0.142(0.006)	**0.108(0.006)**
	6	1.895(0.032)	1.623(0.036)	2.598(0.071)	0.462(0.007)	**0.196(0.007)**	0.197(0.013)

Example 5: This example is used to illustrate the advantages of AGMM methods under the infinite dimensional setting. With a large enough d, e.g. $d = 25$, the data is generated as follows so that Conditions 3 and 4 are satisfied. To be specific, we generate $X_t(u) = \sum_{j=1}^{d} \xi_{tj} \phi_j(u)$ based on $\xi_{tj} = 0.8\xi_{t-1,j} + \epsilon_{tj}$, where $\epsilon_{tj} \sim N(0, j^{-0.75})$. Some specific calculations yield lag-k autocovariance of ξ_{tj} as $\mathrm{Cov}(\xi_{tj}, \xi_{t+k,j}) = \frac{0.8^k \cdot j^{-0.75}}{0.36}$ and eigenvalues of K in equation (9) as $\theta_j = \sum_{k=1}^{L} \mathrm{Cov}(\xi_{tj}, \xi_{t+k,j})^2 = \frac{\sum_{k=1}^{L} 0.8^{2k}}{0.36^2} \cdot j^{-1.5} \asymp j^{-1.5}$ under the orthonormality of $\{\phi_j(\cdot)\}_{j \geqslant 1}$. Hence, Condition 3 is satisfied with $\alpha = 1.5$. Moreover, we set $\tau = 2$ in Condition 4 so that $\tau \geqslant \alpha + 1/2$ is satisfied and hence generate the slope function $\beta_0(\cdot) = \sum_{j=1}^{d} \widetilde{b}_j \phi_j(\cdot)$ with $\widetilde{b}_j = (-1)^{j-1} \cdot 2 \cdot j^{-2}$. The innovations $\{\nu_{tj}\}_{n \times 10}$ are independent $N(0,1)$ variables. The truncated dimension M is chosen so that the top M

eigenvalues explains over 90% of the total variation. Table 8 reports numerical results for all comparison methods under two settings, where $\{\phi_j(\cdot)\}_{j=1}^{d}$ and $\{\zeta_j(\cdot)\}_{j=1}^{10}$ are generated from the corresponding basis functions used in Example 1 and 2, respectively. Again we observe the prominent superiority of two versions of AGMM methods over the competitors with AGMM significantly outperforming Base AGMM.

Table 8: *Example 5*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs. The lowest values are in bold font.

$\{\phi_j\}_{j=1}^{25}, \{\zeta_j\}_{j=1}^{10}$	n	Base CLS	CLS	Base CGMM	Base ALS	Base AGMM	AGMM
Example 1	400	0.972(0.017)	1.068(0.022)	0.913(0.030)	0.708(0.012)	0.582(0.013)	**0.390(0.017)**
	800	0.810(0.011)	0.849(0.012)	0.540(0.018)	0.535(0.008)	0.329(0.008)	**0.200(0.008)**
	1200	0.775(0.009)	0.800(0.009)	0.446(0.017)	0.463(0.006)	0.235(0.005)	**0.156(0.005)**
Example 2	400	0.677(0.012)	0.684(0.015)	0.838(0.026)	0.702(0.012)	0.590(0.014)	**0.376(0.017)**
	800	0.536(0.008)	0.541(0.007)	0.449(0.012)	0.546(0.007)	0.341(0.008)	**0.200(0.007)**
	1200	0.482(0.005)	0.486(0.005)	0.308(0.009)	0.478(0.004)	0.241(0.005)	**0.153(0.005)**

References

Bathia, N., Yao, Q. and Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, **38**, 3352-3386.

Borovskik, Y. V. (1996). *U Statistics in Banach Spaces*. VSP, Netherlands.

Bosq, D. (2000). *Linear Processes in Function Spaces - Theory and Applications*. Springer, New York.

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