Regression with Functional Errors-in-Predictors: A Generalized Method-of-Moments Approach

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

[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 regression is an important topic in functional data analysis. Traditionally, one often assumes that samples of the functional predictor are independent realizations of an underlying stochastic process, and are observed over a grid of points contaminated by independent and identically distributed measurement errors. In practice, however, the dynamical dependence across different curves may exist and the parametric assumption on the measurement error covariance structure could be unrealistic. In this paper, we consider functional linear regression with serially dependent functional predictors, when the contamination of predictors by the measurement error is "genuinely functional" with fully nonparametric covariance structure. Inspired by the fact that the autocovariance operator of observed functional predictors automatically filters out the impact from the unobservable measurement error, we propose a novel autocovariance-based generalized method-of-moments estimate of the slope parameter. The asymptotic properties of the resulting estimators under different functional scenarios are established. We also demonstrate that our proposed method significantly outperforms possible competitors through intensive simulation studies. Finally, the proposed method is applied to a public financial dataset, revealing some interesting findings.

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

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 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 each functional predictor as being either fully observed (Hall and Horowitz, 2007) or incompletely observed, with measurement error, at a grid of time points. In the latter case, the errors associated with distinct observation points are assumed to be uncorrelated (Crambes et al., 2009). Therefore, the resulting measurement error process can not be viewed as a "genuinely functional" component corresponding to a fully nonparametric covariance structure.

In this paper, we consider the more difficult regression problem involving serially dependent functional predictors contaminated by "genuinely functional" measurement errors. 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 measurement 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 conventional least square approach (Hall and Horowitz, 2007) relies on the sample covariance operator of $W_t(u)$, which is not a consistent estimator for

the true covariance operator of $X_t(u)$, thus failing to account for the contamination that can result in substantial estimation bias. Inspired from a simple fact that $\text{Cov}\{W_t(u), W_{t+k}(v)\} = \text{Cov}\{X_t(u), X_{t+k}(v)\}$ 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 parameter. 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 *least squares* (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 "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 operators for $X_t(\cdot)$ and $e_t(\cdot)$, respectively, the corresponding two covariance operators are identifiable (Descary and Panaretos, 2017). 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 operator to get rid of the effect from the "genuinely functional" measurement error. Surprisingly, it turns out that the autocovariance-induced operator in AGMM is identical to the nonnegative operator in Bathia et al. (2010), which is used to assess the dimensionality of $X_t(\cdot)$ based on observed curves $W_t(\cdot)$ in equation (1). We believe that the autocovariance-based idea adopted in AGMM can potentially be applied in many errors-contaminated curve time series modelling problems. Third, the proposed method can be applied to both scalar and functional responses with either finite or infinite dimensional functional predictors. Theoretically we establish relevant convergence rates for our proposed estimators under different model settings. Finally, empirically we illustrate the superiority of AGMM relative to its 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 the AGMM fitting procedures for both scalar and functional responses. In Section 3, we investigate the asymptotic properties of our proposed estimators for the slope parameter under different functional scenarios. Section 4 illustrates the finite sample performance of AGMM through a series of simulation studies and a public financial dataset. We summarize our paper and discuss several possible future works in Section 5. We relegate all the technical proofs to the Appendix.

2 Methodology

2.1 Model setup

In this section, we describe the model setup for the functional regression with dependent and erroneous 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 regression model is of the form

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

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 we wish to estimate.

We assume that the observed functional predictors $W_1(\cdot), \ldots, W_n(\cdot)$ satisfy the contamination model described in equation (1). The existence of the unobservable noise term $e_t(\cdot)$ reflects the fact that the curves of interest, $X_t(\cdot)$, are not directly observed. Instead, they are measured on a grid of points and are contaminated by functional measurement errors, $e_t(\cdot)$, without assuming any parametric structure on the covariance operator of the noise term, denoted by $C_e(u, v) = \mathrm{Cov}\{e_t(u), e_t(v)\}$. This modelling 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.

For an integer $k \geqslant 0$, we assume that the lag-k autocovariance operator of $X_t(\cdot)$, denoted by $C_k(u, v) = \mathrm{Cov}\{X_t(u), X_{t+k}(v)\}$ does not depend on t. In particular, $C_0(u, v)$ reduces

to the covariance operator of $X_t(\cdot)$, which admits the Karhunen-Loéve 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. The eigen-pairs $\{\lambda_j, \phi_j(\cdot)\}_{j=1,2,\ldots}$ satisfy the eigen-decomposition $\int_{\mathcal{U}} C_0(u, v)\phi_j(v)dv = \lambda_j\phi_j(u)$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots$. We say that $X_t(\cdot)$ is d-dimensional if $\lambda_d \neq 0$, and $\lambda_{d+1} = 0$, for some integer $d \geqslant 1$. When d is finite, $\beta_0(\cdot)$ is not identifiable in general and here we define $\beta_0(\cdot)$ by the minimizer of $\int_{\mathcal{U}} \beta_0^2(u)du$ subject to $\text{Cov}\{Y_1, X_1(u)\} = \int_{\mathcal{U}} C_0(u, v)\beta(v)dv, u \in \mathcal{U}$, which implies that $\beta_0(u) = \sum_{j=1}^{d} \lambda_j^{-1}\text{Cov}(Y_1, \xi_{1j})\phi_j(u)$. When $d = \infty$, all the eigenvalues are nonzero and $X_t(\cdot)$ is a truly infinite dimensional functional object. In this case, provided that $\sum_{j=1}^{\infty} \lambda_j^{-2}\{\text{Cov}(Y_1, \xi_{1j})\}^2 < \infty$, $\beta_0(\cdot)$ can be uniquely expressed as $\beta_0(u) = \sum_{j=1}^{\infty} \lambda_j^{-1}\text{Cov}(Y_1, \xi_{1j})\phi_j(u)$. 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

$$g_k^X(\beta, u) = \text{Cov}\big\{Y_t, X_{t+k}(u)\big\} - \int_{\mathcal{U}} \text{Cov}\{X_t(v), X_{t+k}(u)\}\beta(v)dv. \tag{3}$$

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

$$g_k^X(\beta_0, u) \equiv 0 \text{ for any } u \in \mathcal{U} \text{ and } k = 1, \ldots. \tag{4}$$

In particular, the conventional LS approach is based on (4) with $k = 0$. However, this approach is inappropriate when $X_t(\cdot)$ are replaced by the surrogates $W_t(\cdot)$ given the fact that $C_W(u, v) = \text{Cov}\{W_t(u), W_t(v)\} = C_0(u, v) + C_e(u, v)$, 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 discussion on 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. Our method is based on the simple fact that

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

Then after substituting $X_t(\cdot)$ by $W_t(\cdot)$ in (3), 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 (4) 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 $\mathbf{a}(\cdot) = \big(a_1(\cdot), \ldots, a_L(\cdot)\big)^T$. To simplify our further derivation, 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 $u \in \mathcal{U}$,

$$\sum_{k=1}^{L}\left[\int_{\mathcal{U}} C_k(u, z)\mathrm{Cov}\big\{Y_t, W_{t+k}(z)\big\}dz - \int_{\mathcal{U}}\Big\{\int_{\mathcal{U}} C_k(u, z)C_k(v, z)dz\Big\}\beta(v)dv\right] = 0. \tag{5}$$

To ease our presentation, we define

$$R(u) = \sum_{k=1}^{L}\int_{\mathcal{U}} C_k(u, z)\mathrm{Cov}\{Y_t, W_{t+k}(z)\}dz \tag{6}$$

and the nonnegative operator,

$$K(u, v) = \sum_{k=1}^{L}\int_{\mathcal{U}} C_k(u, z)C_k(v, z)dz. \tag{7}$$

Indeed, the operator K was proposed in Bathia et al. (2010) to identify the dimensionality of $X_t(\cdot)$ based on $W_t(\cdot)$ in equation (1). Substituting the relevant terms in (5), $\beta_0(\cdot)$ satisfies the following operator equation

$$R(u) = \int_{\mathcal{U}} K(u,v)\beta(v)dv \text{ for any } u \in \mathcal{U}, \tag{8}$$

which can be understood as a functional extension of the least squares type of population normal equation.

Provided that $X_t(\cdot)$ is d-dimensional, it follows from Proposition 1 of Bathia et al. (2010) that the operator K has exactly d nonzero eigenvalues, $\theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_d$. Let $\psi_1, \ldots, \psi_d$ be the orthonormal eigenfunctions of K with $\int_{\mathcal{U}} K(u,v)\psi_j(v)dv = \theta_j\psi_j(u)$ for $j = 1, \ldots, d$. Then the spectral decomposition of K takes the form of $K(u,v) = \sum_{j=1}^{d} \theta_j\psi_j(u)\psi_j(v)$.

To solve β in (8), one need to take the "inverse" of K. This operator, however, is not generally invertible. To deal with this issue, we use the Moore-Penrose inverse of K, written as K^{-1}. 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 $\widetilde{K} = K\big|\ker(K)^{\perp}$, 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 of (8) in $\ker(K)^{\perp}$ and can take the form of

$$\beta_0(u) = \int_{\mathcal{U}} K^{-1}(u,v)R(v)dv = \sum_{j=1}^{d} \theta_j^{-1}\langle\psi_j, R\rangle\psi_j(u). \tag{9}$$

Provided K is a bounded operator under the infinite dimensional setting ($d = \infty$), 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, for example 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 $\sum_{j=1}^{\infty} \theta_j^{-2}\langle\psi_j, R\rangle^2 < \infty$, an unique solution of (8) exists and is of the form

$$\beta_0(u) = \int_{\mathcal{U}} K^{-1}(u,v)R(v)dv = \sum_{j=1}^{\infty} \theta_j^{-1}\langle\psi_j, R\rangle\psi_j(u). \tag{10}$$

2.3 Estimation procedure

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

We first provide the sample versions of $C_k(u,v)$ and $\text{Cov}\{Y_t, W_{t+k}(u)\}$ for $k = 1, \ldots, L$, which are

$$\widehat{C}_k(u,v) = \frac{1}{n-L}\sum_{t=1}^{n-L} W_t(u)W_{t+k}(v) \quad \text{and} \quad \widehat{\text{Cov}}\{Y_t, W_{t+k}(u)\} = \frac{1}{n-L}\sum_{t=1}^{n-L} Y_t W_{t+k}(u). \tag{11}$$

Combing (6), (7) and (11) gives the the natural estimators for $K(u,v)$ and $R(u)$ as

$$\widehat{K}(u,v) = \sum_{k=1}^{L}\int_{\mathcal{U}} \widehat{C}_k(u,z)\widehat{C}_k(v,z)dz = \frac{1}{(n-L)^2}\sum_{k=1}^{L}\sum_{t=1}^{n-L}\sum_{s=1}^{n-L} W_t(u)W_s(v)\langle W_{t+k}, W_{s+k}\rangle \tag{12}$$

and

$$\widehat{R}(u) = \sum_{k=1}^{L}\int_{\mathcal{U}} \widehat{C}_k(u,z)\widehat{\text{Cov}}\{Y_t, W_{t+k}(z)\}dz = \frac{1}{(n-L)^2}\sum_{k=1}^{L}\sum_{t=1}^{n-L}\sum_{s=1}^{n-L} W_t(u)Y_s\langle W_{t+k}, W_{s+k}\rangle, \tag{13}$$

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

We next perform an eigenanlysis on $\widehat{K}$ and thus obtain the estimated eigen-pairs $\{\widehat{\theta}_j, \widehat{\psi}_j(\cdot)\}$ for $j = 1, 2, \ldots$. When the number of functional observations n is large, the accumulated errors in (12), (13) 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 eigen-pairs $\{(\widehat{\theta}_j, \widehat{\boldsymbol{\delta}}_j)\}_{j=1}^{J}$. 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 details on the selection of J. A similar basis function expansion technique can be applied to produce a smooth estimate $\widehat{R}(\cdot)$. Note that $\widehat{\mathbf{K}}, \widehat{\theta}_j, \widehat{\psi}_j(\cdot), j = 1, \ldots, d$, 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 (9) and (10) 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 $\widehat{\beta}(u)$ of $\beta_0(u)$ is then given by

$$\widehat{\beta}(u) = \sum_{j=1}^{\widehat{d}} \widehat{\theta}_j^{-1} \langle \widehat{\psi}_j, \widehat{R} \rangle \widehat{\psi}_j(u). \tag{14}$$

(ii) When $X_t(\cdot)$ is an infinite dimensional functional object, we take the standard truncation approach by using the leading M eigen-pairs of $\widehat{K}$ to approximate $\beta_0(u)$ in (10). Specifically, we obtain the estimated slope function as

$$\widehat{\beta}(u) = \sum_{j=1}^{M} \widehat{\theta}_j^{-1} \langle \widehat{\psi}_j, \widehat{R} \rangle \widehat{\psi}_j(u). \tag{15}$$

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 $\beta_0(\cdot)$. Specifically, we propose

$$\widehat{\beta}_{\text{ridge}}(u) = \sum_{j=1}^{\widetilde{M}} (\widehat{\theta}_j + \rho_n)^{-1} \langle \widehat{\psi}_j, \widehat{R} \rangle \widehat{\psi}_j(u), \tag{16}$$

where $\widetilde{M}$ is chosen to be reasonably larger than M and ρ_n is a non-negative ridge parameter. See also Hall and Horowitz (2007) for the ridge-type estimator in the classical functional 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 regression takes the form of

$$Y_t(u) = \int_{\mathcal{U}} X_t(v)\gamma_0(u,v)dv + \varepsilon_t(u),\ u \in \mathcal{U},\ t = 1, \ldots, n, \tag{17}$$

where $\gamma_0(u,v)$ is the slope parameter 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 component $X_t(\cdot)$ itself, as described in equation (1) and Section 2.1.

To estimate the slope operator in (17), we develop an AGMM approach analogous to that for the scalar case in Section 2 by solving the normal equation of

$$H(u,v) = \int_{\mathcal{U}} K(u,w)\gamma(w,v)dw \text{ for any } v \in \mathcal{U}, \tag{18}$$

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

$$\widehat{H}(u,v) = \frac{1}{(n-L)^2} \sum_{k=1}^{L} \sum_{t=1}^{n-L} \sum_{s=1}^{n-L} W_t(u)Y_s(v)\langle W_{t+k}, W_{s+k}\rangle. \tag{19}$$

Accordingly, we can provide the estimate $\widehat{\gamma}$ of γ_0 under two functional scenarios involving $d < \infty$ and $d = \infty$. (i) Under the finite dimensional setting ($d < \infty$), $\gamma_0(u,v)$ is the unique solution of (18) in $\ker(K)^{\perp}$ and can be represented as

$$\gamma_0(u,v) = \int_{\mathcal{U}} K^{-1}(u,w)H(w,v)dw = \sum_{j=1}^{d} \theta_j^{-1}\langle \psi_j, H(\cdot,v)\rangle \psi_j(u). \tag{20}$$

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

$$\widehat{\gamma}(u,v) = \sum_{j=1}^{\hat{d}} \widehat{\theta}_j^{-1}\widehat{\psi}_j(u)\langle \widehat{\psi}_j, \widehat{H}(\cdot,v)\rangle. \tag{21}$$

(ii) Under the infinite dimensional setting ($d = \infty$), if we assume the boundedness of the composite operator $\int_{\mathcal{U}} K^{-1}(u,w)H(w,v)dw$ in the L_2 sense, then the solution of (18) uniquely exists. Approximating the infinite dimensional $\gamma_0(u,v)$ in (20) by the first M components and substituting the relevant terms by their estimated values, we can obtain the estimated slope operator as

$$\widehat{\gamma}(u,v) = \sum_{j=1}^{M} \widehat{\theta}_j^{-1}\widehat{\psi}_j(u)\langle \widehat{\psi}_j, \widehat{H}(\cdot,v)\rangle. \tag{22}$$

2.5 Selection of tuning parameters

Implementing AGMM requires choosing L (selected lag length in (5)), M (truncated dimension in (15) 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 the 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 to choose M under various situations include the bootstrap test (Bathia et al., 2010) and the eigen-ratio-based estimator (Lam et al., 2011; Lam and Yao, 2012). 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 (12), 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 $\widehat{\boldsymbol{\Delta}}^{(-g)} = \big(\widehat{\boldsymbol{\delta}}_1^{(-g)}, \ldots, \widehat{\boldsymbol{\delta}}_d^{(-g)}\big) \in \mathbb{R}^{J\times d}$, formed by 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 $\widehat{\theta}_l^{(g)}, l = 1.\ldots, d$ be the diagonal elements of $(\widehat{\boldsymbol{\delta}}_1^{(-g)})^{\tau}\widehat{\mathbf{K}}^{(g)}\widehat{\boldsymbol{\delta}}_1^{(-g)}$.

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

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

Given the time break on the training observations, the autocovariance assumption is jeopardized by $L = 5$ misutilized lagged terms. However, this effect on the estimate $\widehat{K}$ is negligible especially when n is sufficiently large, hence our proposed CV approach can still be applied in practice. See also Bergmeir et al. (2018) for the discussion on 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 regressions.

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

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

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

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) The eigenfunctions $\{\psi_j(\cdot)\}_{j=1}^{\infty}$ are continuous on the compact set $\mathcal{U}$ and satisfy $\sup_{j\geqslant 1}\sup_{u\in\mathcal{U}}|\psi_j(u)| = O(1)$.*

Condition 4 *When $d = \infty$, $\beta_0(u) = \sum_{j=1}^{\infty} b_j\psi_j(u)$ and there exist some positive constants $\tau \geqslant 3/2$ 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 $\theta_j \geqslant cj^{-\alpha}$ as $\theta_j = \sum_{k=j}^{d}(\theta_k - \theta_{k+1}) \geqslant c\sum_{k=j}^{d} k^{-\alpha-1}$, and can be used to derive the convergence rates of estimated eigen-pairs. See also Hall and Horowitz (2007) and Qiao et al. (2017). Under the $d = \infty$ setting, Condition 4 restricts the true slope function based on the expansion of $\beta_0(\cdot)$ using the eigenfunctions of K. The parameter τ determines the decay rate of slope basis coefficients, $\{b_j\}_{j=1}^{\infty}$. The assumption $\tau \geqslant 3/2$ can be interpreted as requiring that β_0 be sufficiently smooth relative to K, the smoothness of which can be implied by $\theta_j \geqslant cj^{-\alpha}$ with $\alpha > 1$ from Condition 3(ii). See Hall and Horowitz (2007) for an analogous condition in functional regression.

Before presenting Theorems 1 and 2, which correspond to the asymptotic results for the functional regression with scalar and functional responses respectively, we first solidify some notation. For any function f, define $||f|| = \sqrt{\langle f, f\rangle}$. We denote by $||A||_{\mathcal{S}}$ the Hilbert-Schmidt norm for any operator 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 (14) under the $d < \infty$ setting, 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

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

We provide remarks on different convergence rates under two functional scenarios. First, when d is fixed, the standard parametric root-n rate is achieved. Second, 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 $O_P\{n^{-(2\tau-1)/(\alpha+2\tau)}\}$ developed in Hall and Horowitz (2007), which considers independent observations of the functional predictor without measurement errors. In fact, we tackle a more difficult functional regression problem, where extra complications come from the serial dependence and functional measurement error structure. 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 regression, we first list two regularity conditions in 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 $\tau \geqslant 3/2$ 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 Empirical studies

4.1 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 $\{\xi_{tj}\}_{t=1}^{T}$ 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 operator $\widehat{C}_W$, which converts the functional linear regression to the multiple linear regression, and then implement either LS or GMM. The 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-induced operator $\widehat{K}$ 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

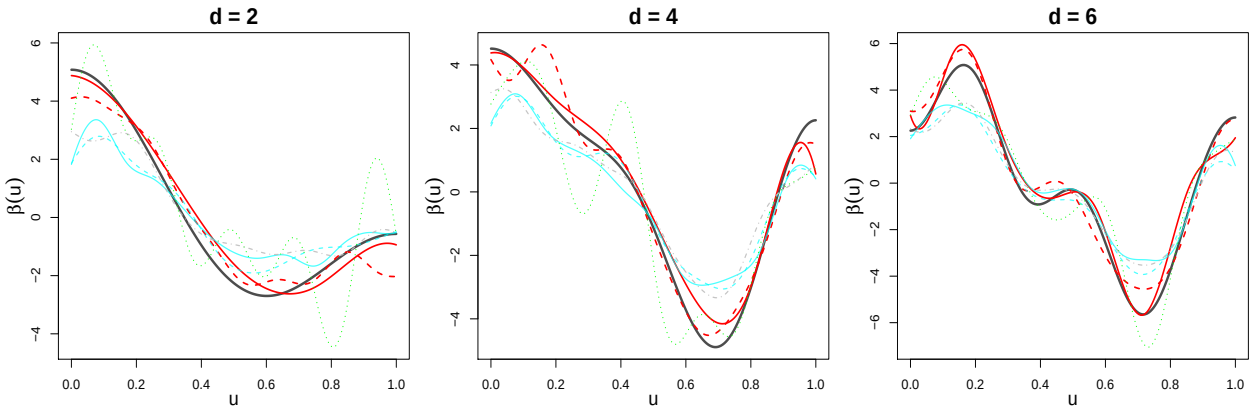

Figure 1: *Example 1 with* $n = 800$ *and* $d = 2, 4, 6$*: Comparison of true* $\beta(\cdot)$ *functions (black solid) with median estimates over 100 simulation runs for AGMM (red solid), Base AGMM (red dashed), CLS (cyan solid), Base CLS (cyan dashed), Base CGMM (green dotted) and Base ALS (gray dash-dotted).*

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.

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 measurement error; (iii) the estimate $\widehat{C}_W$ in CGMM does not consider the measurement error, while $\widehat{K}$ in AGMM would suffer

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)

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 3 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.

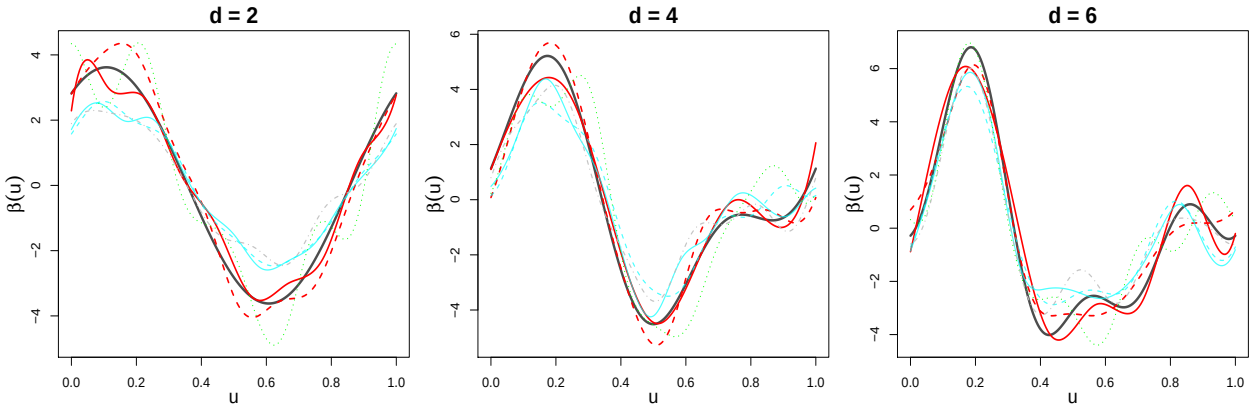

Figure 2: *Example 2 with* $n = 800$ *and* $d = 2, 4, 6$*: Comparison of true* $\beta(\cdot)$ *functions (black solid) with median estimates over 100 simulation runs for AGMM (red solid), Base AGMM (red dashed), CLS (cyan solid), Base CLS (cyan dashed), Base CGMM (green dotted) and Base ALS (gray dash-dotted).*

4.2 Real data analysis

In this section, we illustrate the proposed AGMM using a public financial dataset. Let $P_t(u_j), t = 1, \ldots, n, j = 1, \ldots, T$ be the price of a financial asset at time u_j on the t-th trading day. Denote the *cumulative intraday return* (CIDR) trajectory, in percentage, by $r_t(u_j) = 100\big[\log\{P_t(u_j)\} - \log\{P_t(u_1)\}\big]$, where u_1 is the opening time of the trading day (Horvath et al., 2014). The dataset we consider was downloaded from `https://wrds-web.wharton.upenn.edu/wrds` and consists of one-minute resolution prices of Standard & Poor's 100 index and inclusive stocks from $n = 251$ trading days in year 2017. The trading period (9:30-16:00) with $T = 390$ minutes is rescaled onto $\mathcal{U} = [0, 1]$. We first obtain the smoothed

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)

CIDR curves on the Standard & Poor's 100 index using the standard kernel method, $\widetilde{r}_{m,t}(u) = (Th_t)^{-1}\sum_{j=1}^{T} K_{h_t}(r_{m,t}(u_j) - u)$, where $K_h(u)$ is a kernel function with bandwidth h. Let $\widehat{\sigma}_t$ be the sample standard deviation of $r_{m,t}(u_j)$ for $j = 1, \ldots, T$. We use a Gaussian kernel with the optimal bandwidth $\widehat{h}_t = 1.06\widehat{\sigma}_t T^{-1/5}$ (Silverman, 1999).

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

$$r_t = \alpha + \int X_t(u)\beta(u)du + \varepsilon_t, \quad \widetilde{r}_{m,t}(u) = X_t(u) + e_t(u), \tag{23}$$

where $X_t(\cdot)$ and $e_t(\cdot)$ represent the signal and error components in $\widetilde{r}_{m,t}(\cdot)$, respectively, and $r_t = 100\big[\log\{P_t(u_T)\} - \log\{P_{t-1}(u_T)\}\big]$ is the return for a specific stock on the t-th trading day. Note that the slope parameter in the classical capital asset pricing models explains how strongly an asset return depends on the market portfolio. Analogously, $\beta(\cdot)$ in (23) can

be understood as the functional sensitivity measure of an asset return to the market CIDR trajectory.

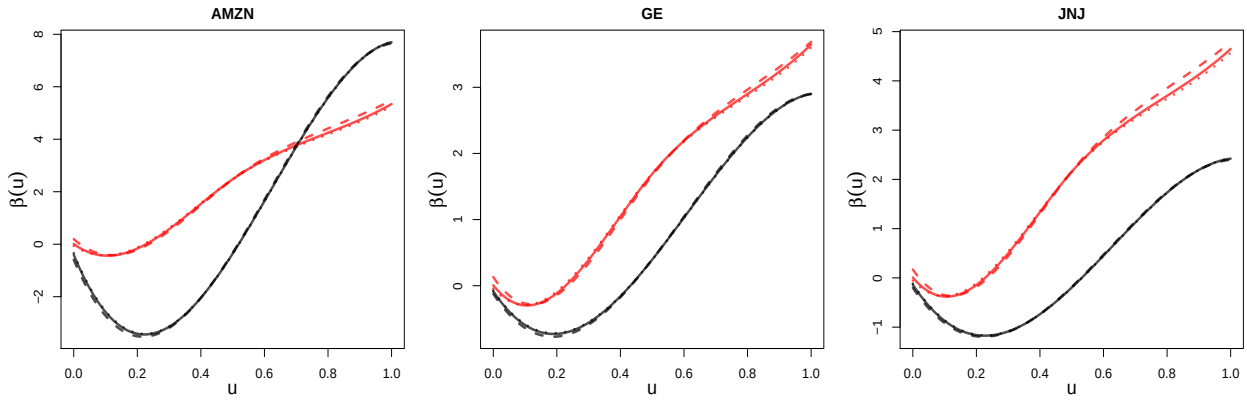

Figure 3: *Estimated $\beta(\cdot)$ curves for AGMM (red) and CLS (black) using $h_t = \widehat{h}_t$ (solid lines), $0.5\widehat{h}_t$ (dotted lines) and $2\widehat{h}_t$ (dashed lines).*

Figure 3 plots the estimated $\beta(\cdot)$ functions using both CLS and AGMM for three large-cap-sector stocks, Amazon (ticker AMZN), General Electronic (ticker GE) and Johnson & Johnson (ticker JNJ). To identify the finite dimensionality of $\widetilde{r}_{m,t}(\cdot)$, we apply the bootstrap test and the eigen-ratio-based estimator (Lam and Yao, 2012). Both approaches suggest to take $\widehat{d} = 1$. We observe a few apparent patterns in Figure 3. 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 estimates are insensitive to the choice of bandwidth and the shapes of the estimated $\beta(\cdot)$ functions by either CLS or AGMM are quite similar across the three stocks.

5 Discussion

We conclude our paper with several remarks. First, in comparison with the classical functional regression setting, we study a more difficult problem by relaxing the critical independence assumption and allowing functional predictors to be corrupted by "genuinely functional" measurement errors. Second, to address the problem we consider, one can possibly adopt the dimension reduction approach for curve time series (Bathia et al., 2010), which

transfers the functional linear regression to the multiple linear regression, however the extra uncertainty from the measurement error would still stop us from using the LS approach while the deficiency of ALS are demonstrated by the simulation studies. Instead, AGMM can successfully solve this issue by using the autocovariance to remove the part due to the noise term. Moreover, the AGMM approach is closely connected to the work of Bathia et al. (2010). In particular, the operator K proposed under the GMM framework is exactly the same as the nonnegative operator in Bathia et al. (2010) based on the same contamination model as expressed in equation (1).

We identify several potential directions for future research. First, we can consider extending the current regression model to the multivariate or even high dimensional setting involving p possibly erroneous functional predictors, where p can be very large. Under the independence and large p, small n, setting, some concentration inequalities based on the covariance structure are established in Qiao et al. (2017). It is of great interest to develop the relevant concentration bounds for high dimensional curve time series under our proposed autocovariance-based framework, which would provide a powerful tool to derive the non-asymptotic upper bounds. The second potential extension concerns the functional singular component analysis (FSCA) (Yang et al., 2011) to model a pair of erroneous curve time series. One possible way to tackle this type of bivariate data is to perform FSCA on some autocovairance-based operator, where the impact from the measurement error can be eliminated. It is worth noting that some functional relationships such as function-on-function regression might also be represented under a FSCA framework, see Cho et al. (2013) for details. Then an analogous autocovariance-based GMM approach could possibly be applied. Third, the convergence rate in Theorem 1(ii) is slightly slower than the one in Hall and Horowitz (2007). It is of great interest to either prove the optimality of our rate or develop the optimal minimax rate under the setting we consider. These topics are beyond the focus of this paper and will be pursued elsewhere.

A Appendix

Appendices A.1 and A.2 contain all the technical proofs.

A.1 Proof of Theorem 1

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 $\delta_j \leqslant C j^{\alpha+1}$ with some positive constant C. This completes our proof for part (iii).

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.

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

Proof. By the definitions of C_k and (6), 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 $||E\{W_t(u)W_{t+k}(v)\}||_{\mathcal{S}}^2$

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

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 $||R|| = O(1)$.

A.1.1 Proof of Theorem 1 (i)

First we provide Lemma 4 to show the consistency of $\widehat{d}$ to d when $d < \infty$.

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.

Define $\check{K}(u,v) = \sum_{j=1}^{d} \widehat{\theta}_j \widehat{\psi}_j(u) \widehat{\psi}_j(v)$ and $K^{-1}(u,v) = \sum_{j=1}^{d} \theta_j^{-1} \psi_j(u) \psi_j(v)$. We have the following result.

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

Proof. Observe that

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

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

$$
||\check{K}^{-1} - K^{-1}||_{\mathcal{S}} \leqslant \sum_{j=1}^{d} \widehat{\theta}_j^{-1} \theta_j^{-1} |\widehat{\theta}_j - \theta_j| + 2 \sum_{j=1}^{d} \theta_j^{-1} ||\widehat{\psi}_j - \psi_j||. \tag{24}
$$

When d is fixed, the smallest eigenvalue θ_d is bounded away from zero. It follows from Lemma 1 (i),(ii) and (24) that there exists some positive constant C such that $||\check{K}^{-1} - K^{-1}||_{\mathcal{S}} \leqslant C(\theta_d^{-2} + \theta_d^{-1}) n^{-1/2}$, 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.

Now we organize our proof for part (i) of Theorem 1, i.e. the case when $d < \infty$. Let $\widetilde{\beta}(u) = \int_{\mathcal{U}} \check{K}^{-1}(u,v)\widehat{R}(v)dv$. For a large $\delta > 0$, by Lemma 4, we have

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

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

$$||\widetilde{\beta}-\beta_0|| \leqslant ||\check{K}^{-1}-K^{-1}||_{\mathcal{S}}||\widehat{R}|| + ||K^{-1}||_{\mathcal{S}}||\widehat{R}-R||. \tag{25}$$

Then it follows from Lemmas 2, 3 and 5 that $||\widetilde{\beta}-\beta_0|| = O_P(n^{-1/2})$. This completes our proof for part (i) of Theorem 1.

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

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

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

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

$$\widehat{\psi}_j - \psi_j = \sum_{k:k\neq j} (\widehat{\theta}_j - \theta_k)^{-1}\psi_k\langle\widehat{\psi}_j, \langle\widehat{K}-K, \psi_k\rangle\rangle + \psi_j\langle\widehat{\psi}_j - \psi_j, \psi_j\rangle. \tag{26}$$

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

Now we are ready to prove Theorem 1(ii) under the $d = \infty$ setting. 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

$$||\widehat{\beta}-\beta_0||^2 \leqslant ||\widehat{\beta}-\beta_M||^2 + ||\beta_M-\beta_0||^2. \tag{27}$$

By (10) 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

$$||\beta_M - \beta_0||^2 = \sum_{j=M+1}^{\infty} b_j^2 \leqslant C \sum_{j=M+1}^{\infty} j^{-2\tau} = O(M^{-2\tau+1}). \tag{28}$$

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

$$\begin{aligned}\|\widehat{\beta} - \beta_M\|^2 &\leqslant 3\sum_{j=1}^{M} \big(\widehat{\theta}_j^{-1} - \theta_j^{-1}\big)^2\langle\psi_j, R\rangle^2 + 3\sum_{j=1}^{M} \widehat{\theta}_j^{-2}\big(\langle\widehat{\psi}_j, \widehat{R}\rangle - \langle\psi_j, R\rangle\big)^2 \\ &\quad + 3M\sum_{j=1}^{M} \theta_j^{-2}\langle\psi_j, R\rangle^2\|\widehat{\psi}_j - \psi_j\|^2 \\ &= 3I_{n1} + 3I_{n2} + 3I_{n3}.\end{aligned} \tag{29}$$

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

$$I_{n1} = O_P(n^{-1}) \cdot \Big(\sum_{j=1}^{M} j^{2\alpha-2\tau}\Big) = O_P(n^{-1}) \cdot \big(M + M^{2\alpha-2\tau+1}\big) = o_P\big(n^{-1}M^{2\alpha+1}\big). \tag{30}$$

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

$$I_{n3} \leqslant M\sum_{j=1}^{M} b_j^2\|\widehat{\psi}_j - \psi_j\|^2 = O_P\big(n^{-1}M^2 + n^{-1}M^{2\alpha-2\tau+4}\big) = O_P\Big(n^{-1}M^{2\alpha+1}\Big), \tag{31}$$

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

$$\begin{aligned} I_{n2} &\leqslant 4\sum_{j=1}^{M} \theta_j^{-2}\big(\langle\widehat{\psi}_j, \widehat{R}\rangle - \langle\psi_j, R\rangle\big)^2 \\ &\leqslant 12\sum_{j=1}^{M} \theta_j^{-2}\Big(\langle\widehat{\psi}_j - \psi_j, R\rangle^2 + \langle\psi_j, \widehat{R} - R\rangle^2 + \langle\widehat{\psi}_j - \psi_j, \widehat{R} - R\rangle^2\Big) \\ &\leqslant 12\sum_{j=1}^{M} \theta_j^{-2}\Big(\langle\widehat{\psi}_j - \psi_j, R\rangle^2 + \|\widehat{R} - R\|^2 + \|\widehat{\psi}_j - \psi_j\|^2\|\widehat{R} - R\|^2\Big), \end{aligned} \tag{32}$$

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

$$\sum_{j=1}^{M} \theta_j^{-2} R_{j2}^2 = O_P(n^{-1}) \cdot \Big(\sum_{j=1}^{M} j^{-2\tau+2\alpha+2}\Big) = o_P\big(n^{-1}M^{2\alpha+1}\big). \tag{33}$$

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

$$\sup_{j\geqslant 1} \sum_{k:k\neq j} \theta_k^2 b_k^2 (\theta_j - \theta_k)^{-2} = O(1), \tag{34}$$

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

$$\begin{aligned} \sum_{j=1}^{M} \theta_j^{-2} R_{j1}^2 &\leqslant 4\sum_{j=1}^{M} \theta_j^{-2} \sum_{k:k\neq j} \theta_k^2 b_k^2 (\theta_j - \theta_k)^{-2} \|\widehat{K} - K\|_{\mathcal{S}}^2 \\ &= O_P(n^{-1}) \cdot \sum_{j=1}^{M} \theta_j^{-2} = O_P(n^{-1}M^{2\alpha+1}). \end{aligned} \tag{35}$$

We turn to prove (34) 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,

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

uniformly in j. Then (34) follows.

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

$$\sum_{j=1}^{M} \theta_j^{-2} \|\widehat{R} - R\|^2 = O_P(n^{-1} M^{2\alpha+1}) \text{ and } \sum_{j=1}^{M} \theta_j^{-2} \|\widehat{\psi}_j - \psi_j\|^2 \|\widehat{R} - R\|^2 = O_P(n^{-2} M^{4\alpha+3}). \quad (36)$$

Combing the results in (32)–(33) and (35)–(36), we have

$$I_{n2} = O_P\Big(n^{-2} M^{4\alpha+3} + n^{-1} M^{2\alpha+1}\Big). \quad (37)$$

Combining the results in (27),(28) and (37) and choosing $M \asymp n^{1/(2\alpha+2\tau)}$, we obtain that

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

which completes the proof.

A.2 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

$$\|\widehat{H} - H\|_{\mathcal{S}} = O_P(n^{-1/2}) \text{ and } \|H\|_{\mathcal{S}} = O(1). \quad (38)$$

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

$$||\widetilde{\gamma} - \gamma_0||_{\mathcal{S}} \leqslant ||\check{K}^{-1} - K^{-1}||_{\mathcal{S}} ||\widehat{H}||_{\mathcal{S}} + ||K^{-1}||_{\mathcal{S}} ||\widehat{H} - H||_{\mathcal{S}}. \quad (39)$$

It follows from Lemma 5 and (38) 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

$$||\widehat{\gamma}-\gamma_0||_{\mathcal{S}}^2\leqslant||\widehat{\gamma}-\gamma_M||_{\mathcal{S}}^2+||\gamma_M-\gamma_0||_{\mathcal{S}}^2. \tag{40}$$

It follows from Condition 6 and some specific calculations that

$$\begin{aligned}\|\gamma_M-\gamma_0\|_{\mathcal{S}}^2 &= \|\sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty}b_{j\ell}\psi_j(u)\psi_\ell(v)\|_{\mathcal{S}}^2\\ &= \sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty}b_{j\ell}^2\leqslant C\sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty}(j+\ell)^{-2\tau-1}=O(M^{-2\tau+1}).\end{aligned} \tag{41}$$

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

$$||\widehat{\gamma}-\gamma_M||_{\mathcal{S}}^2=O_P(M^{4\alpha+3}n^{-2}+M^{2\alpha+1}n^{-1}). \tag{42}$$

Combing the results in (40)–(42) 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.

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.

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

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

Campbell, J., Lo, A. W. and A.C., M. (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. (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. (2017). Functional data analysis by matrix completion, *To appear in the Annals of Statistics* .

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.

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. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors, *The Annals of Statistics* **40**: 694–726.

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. (2017). Functional graphical models, *Journal of the American Statistical Association* **in press**.

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

Silverman, B. (1999). *Density estimation for Statistics and Data Analysis (2nd ed.)*, Chapman and Hall, London.

Yang, W., Mueller, H. G. and Stadtmueller, U. (2011). Functional singular component analysis, *Journal of the Royal Statistical Society: Series B* **73**: 303–324.

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

Supplementary Material to " Regression with Functional Errors-in-Predictors: A Generalized Method-of-Moments Approach"

Xinghao Qiao, Cheng Chen, and Shaojun Guo

This supplementary material contains additional simulation results supporting Section 4.

B Additional simulation results

For Example 2, Table 3 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 4. Under this setting, we can observe that both approaches are capable of correctly identifying the d signal components.

Table 3: 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	0.25	0.13	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 4: 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.

	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**1.73**	**1.19**								
d=2	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
	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**2.50**	**1.73**	**1.38**	**1.19**						
d=4	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
	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**3.00**	**2.16**	**1.73**	**1.47**	**1.30**	**1.19**				
d=6	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