Functional time series prediction under partial observation of the future curve

Shuhao Jiao *1, Alexander Aue †2, and Hernando Ombao ‡3

[1] *Statistics Program, KAUST, Saudi Arabia*

[2] *Department of Statistics, UC Davis, CA, USA*

[3] *Statistics Program, KAUST, Saudi Arabia*

Abstract

This paper tackles one of the most fundamental goals in functional time series analysis which is to provide reliable predictions for functions. Existing functional time series methods seek to predict a complete future functional observation based on a set of observed complete trajectories. The problem of interest discussed here is how to advance prediction methodology to cases where partial information on the next trajectory is available, with the aim of improving prediction accuracy. To solve this problem, we propose a new method "partial functional prediction (PFP)". The proposed method combines "next-interval" prediction and fully functional regression prediction, so that the partially observed part of the trajectory can aid in producing a better prediction for the unobserved part of

* shuhao.jiao@kaust.edu.sa

† aaue@ucdavis.edu

‡ hernando.ombao@kaust.edu.sa

the future curve. In PFP, we include automatic selection criterion for tuning parameters based on minimizing the prediction error. Simulations indicate that the proposed method can outperform existing methods with respect to mean-square prediction error and its practical utility is illustrated in an analysis of environmental and traffic flow data.

Keywords: Dimension reduction, Functional principal component analysis, Final prediction error, Functional time series prediction, Intra-day fully functional linear regression model, Long-term and short-term dynamics, Updating prediction.

1 Introduction

Functional data is collected in many sociological, environmental, biological and clinical research. Of prime interest in this paper is to analyze the daily trajectories of the pollutant PM10 (which are fine particulate matter with diameter less than 10 micrometers) in Graz, Austria, which is displayed in Figure 1. The task in this paper is to develop a new method that can predict the trajectory on Sunday using all the past daily trajectories plus the partially observed trajectory on Sunday. We first review some of the existing approaches to analyzing functional time series data. As noted, functional data are often collected over many natural consecutive time intervals. For the PM10 data discussed above, the dataset consists of many daily curves. One of the interesting aspects of functional time series is that the many trajectories may share similar behavior. The collected functions may be described by a time series $(Y_k \colon k \in \mathbb{Z})$, $\mathbb{Z}$ denoting the integers, with observations in the sequence being random functions $Y_k(t)$ for t taking values in some domain $\mathcal{U}$, here taken to be the unit interval $[0, 1]$. The object $(Y_k \colon k \in \mathbb{Z})$ will be referred to as a functional time series. Interest in this new method arises from the consideration of the dynamic features of functional time series

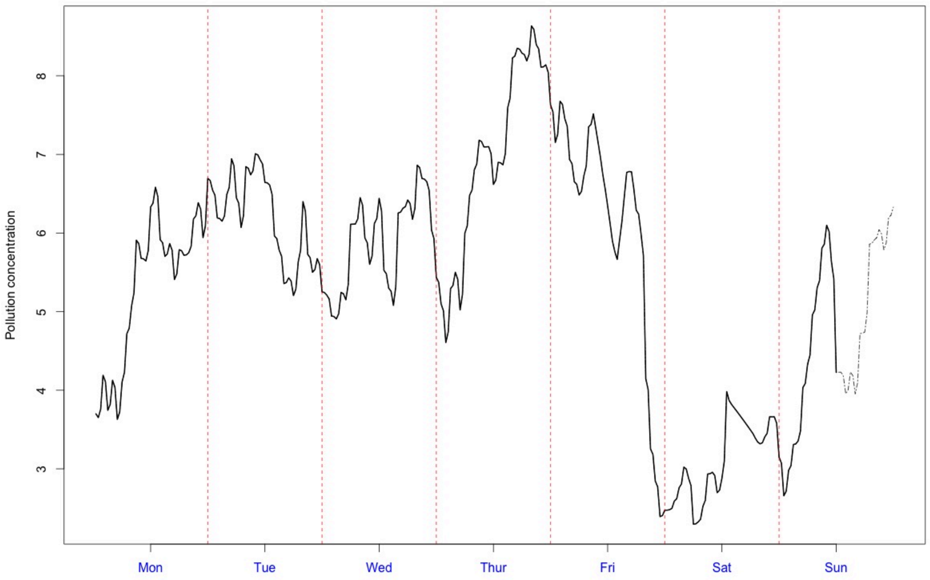

Figure 1: Plot of daily trajectories of the PM10 concentration over one week. The dotted grey line represents the unobserved part of Sunday's trajectory.

data.

Complete curve prediction has been discussed in recent decades. However, these existing methods are not tailored for the specific case where there is partially observed data that is available for predicting a new trajectory. The existing methods focus often on the Functional AutoRegressive model of order p, FAR(p), model. Bosq (2000) derived one-step ahead predictors based on a functional form of the Yule–Walker equations for FAR(1) processes. Besse, Cardot and Stephenson (2000) proposed non-parametric kernel predictors. Antoniadis and Sapatinas (2003) studied FAR(1) curve prediction based on linear wavelet methods. Kargin and Onatski (2008) introduced the predictive factor method, which seeks to replace functional principal components with directions most relevant for predictions. Didericksen, Kokoszka and Zhang (2012) evaluated several competing prediction models in a comparative simulation study, finding Bosq's (2000) method to have the best overall prediction performance based on mean squared error and averaged distance. Aue, Dubart Norinho and Hörmann (2015) proposed a method that deals with functional time series prediction in a multivariate way, together with a final prediction error criterion to select the order of FAR process and the dimension of the auxiliary VAR model. Existing full-curve prediction method for functional time series, only incorporate the dynamics across functions, but not the intra-function information. Thus partial observation is not utilized to improve the prediction. To overcome this limitation, we incorporate the information both across and within curves.

In contrast to complete curve prediction, PFP aims to give predictions based on a partially observed trajectory. Fully functionally regression method has been considered in providing updated time series prediction in Chiou (2012), who proposed a functional mixture method for predicting traffic flow. The proposed method is a combination of fully functional regression with functional clustering and discrimination. Shang (2017) also considered the fully functional regression method, together with moving block method, to update functional time series predictions. However, these papers have some limitations: Moving block method does not incorporate the intra-curve information, and fully functional regression does not take the cross-curve information into account, which would be a problem when there existed strong cross-curve correlation. More details will be discussed in this paper.

Our prediction method uses all available data – both complete trajectories and partial trajectories. Compared with the complete curve prediction, PFP adds flexibility, since it can update the prediction according to different times of day, and the prediction error over the forecasting time interval should then be smaller. The ability to update will have advantages including reduced prediction error. In practical terms, the decay of the eigenvalues will constrain the starting part of the predicted trajectory to be "close" to the ending part of the partially observed data.

The proposed prediction algorithm is a stepwise procedure and can be summarized as follows. For smoothed trajectories, we decompose the observations into two parts:

$$Y_k(t) = S_k(t) + \epsilon_k(t), \qquad k = 1, 2, \dots,$$

where $S_k(t)$ is the signal function, and $\epsilon_k(t)$ is the *i.i.d* innovation function. For $\tau \in [0, 1]$, assume that the sub-function over interval $[0, \tau]$, denoted by $Y_{n+1}|_{[0,\tau]}$, has already been observed and that the goal is to predict $Y_{n+1}|_{(\tau,1]}$. To do this, we first use functional time series methodology to calculate the fitted functions $\hat{Y}_k(t)$ for all $t \in [0, 1]$, and obtain the residual functions $\hat{\epsilon}_k(t) = Y_k(t) - \hat{Y}_k(t)$. We then separate the residual functions into two segments $\hat{\epsilon}_k|_{[0,\tau]}$ and $\hat{\epsilon}_k|_{(\tau,1]}$ at the current time τ, and fully functionally regress $\hat{\epsilon}_k|_{(\tau,1]}$ on $\hat{\epsilon}_k|_{[0,\tau]}$. The fitted function $\hat{\hat{\epsilon}}_{n+1}|_{(\tau,1]}$ is then used to update the prediction of the unobserved part of the innovation function of the current curve. The final prediction

$$\hat{Y}^u_{n+1}|_{(\tau,1]} = \hat{Y}_{n+1}|_{(\tau,1]} + \hat{\hat{\epsilon}}_{n+1}|_{(\tau,1]}$$

is proposed to be the summation of predictions at each step, where $\hat{Y}_{n+1}$ is the full-curve prediction.

In the noisy case, we further decompose the observations into three parts:

$$Y_k(t) = S_k(t) + \epsilon_k(t) + e_k(t), \qquad k = 1, \dots, n,$$

In addition to the two aforementioned stages, we propose one more step to extract the time series information in the random error $e_k(t)$ on the first two moments, which represents short-term dynamics. In this article, we will discuss the following: (1) How well does PFP perform, compared with “next interval” prediction method and fully functional regression method? (2) How to select the tuning parameters? (3) How to adjust the method such that it will still produce decent and reasonable prediction for noisy data?

The remainder of the paper is organized as follows. In Section 2, we describe the functional time series prediction methodology proposed by Aue et al. (2015), and discuss the fully functional linear model, and its application to intra-day prediction. We also propose a data-driven criterion of parameter selection for the prediction by fully functional regression model. Section 3 gives the prediction algorithm for both smooth and noisy functional time series. Section 4 shows simulation results, including the prediction MSEs of various methods, the result of order and dimension selection, and nonparametric bootstrap prediction intervals. Real data analyses on PM10 concentration curves and traffic flow trajectories are shown in Section 5.

2 Functional Autoregressive Model and Fully Functional Regression Model

The two popular classes of models for analyzing functional data are functional autoregressive models (FAR) and fully functional regression models. FAR models are used for analyzing a series of correlated functional data (a time series of curves, or a time series of functions). Fully functional regression models are utilized to find the linear relation between two sets of functions. In this paper, we develop a prediction method using principles that are inspired by these two approaches. Before giving the introduction of these two models, we first introduce the following assumptions.

2.1 Preliminaries

Let $(Y_k \colon k \in \mathbb{Z})$ be an arbitrary stationary functional time series satisfying the following assumptions:

(A.1) All random functions are defined on some common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The notation $Y \in L_H^p = L_H^p(\Omega, \mathcal{A}, \mathbb{P})$ is used to indicate that, for some $p > 0$, $E[\|Y\|^p] < \infty$. We assume the observations Y_k are elements of the Hilbert space $H = L^2([0,1])$ equipped with the inner product $\langle x, y \rangle = \int_0^1 x(t)y(t)dt$. Each Y_k is a square integrable function satisfying $\|Y_k\|^2 = \int_0^1 Y^2(t)dt < \infty$.

(A.2) Any $Y \in L_H^1$ possesses a mean curve $\mu = (E[Y(t)] \colon t \in [0,1])$, and any $Y \in L_H^2$ possess a covariance operator C, defined by $C(x) = E[\langle Y - \mu, x\rangle (Y - \mu)]$, equivalently, $C(x)(t) = \int_0^1 c(t,s)x(s)ds$, $c(t,s) = \operatorname{cov}(Y(t), Y(s))$. By spectral decomposition, we have the following expression of C,

$$C(x) = \sum_{j=1}^{\infty} \lambda_j \langle v_j, x \rangle v_j,$$

where $(\lambda_j \colon j \in \mathbb{N})$ are the eigenvalues (in strictly descending order) and $(v_j \colon j \in \mathbb{N})$ the corresponding normalized eigenfunctions (fPC), so that $C(v_j) = \lambda_j v_j$ and $\|v_j\| = 1$.

(A.3) The $(v_j \colon j \in \mathbb{N})$ form an orthonormal basis of $L^2([0,1])$. Then by the statement of Karhunen–Loève theorem, Y_k allows for the representation

$$Y_k = \mu + \sum_{j=1}^{\infty} \langle Y_k - \mu, v_j \rangle v_j, \qquad k \in \mathbb{Z}.$$

The coefficients $\langle Y_k - \mu, v_j \rangle$ in this expansion are called the fPC scores of Y_k. Suppose now that we have observed $Y_1, \ldots, Y_n$. In practice μ as well as C and its spectral decomposition should be unknown and need to be estimated from the sample. We estimate μ pointwise by

$$\hat{\mu}_n(t) = \frac{1}{n} \sum_{k=1}^{n} Y_k(t), \qquad t \in [0,1],$$

and the covariance operator by

$$\hat{C}_n(x) = \frac{1}{n}\sum_{k=1}^{n}\langle Y_k - \hat{\mu}_n, x\rangle (Y_k - \hat{\mu}_n), \qquad x \in H.$$

Remark. In (A.1), all functions are defined in the same probability space so that we can extract the common information of the functional time series. Each function is constrained with proper values so that the covariance operator is well defined. The eigenfunctions $(\nu_j \colon j \in \mathbb{N})$ in Assumption (A.2) form a series of orthonormal basis for the space H, and comparing with other orthonormal basis, the eigenfunctions will give the best approximation of functions with the same number of basis. Assumption (A.3) presents the Karhunen-Loéve expansion which is fundamental for the prediction of functional time series. More specifically, in PFP, the prediction of functional time series is done in the subspace spanned by the first few eigenfunctions, and the predicted function is represented by truncated Karhunen-Loéve expansion.

2.2 Multivariate technique of predicting FAR(p) process

There are existing methods for prediction of functional time series. Among them, Aue et al. (2015) proposed a dimension-reduction method for prediction of stationary functional time series which can be easily implemented and provides competitive prediction results. To determine the order of the FAR model and the dimension of the auxiliary projected eigenspace, the functional final prediction error criterion was proposed. The FAR(p) process is defined by the stochastic recursion

$$Y_k - \mu = \sum_{j=1}^{p} \Phi_j (Y_{k-j} - \mu) + \epsilon_k,$$

where $(\epsilon_k \colon k \in \mathbb{Z})$ are centered, independent and identically distributed innovations in L_H^2 and $\Phi_j \colon H \to H$ are bounded linear operators. The FAR(p) process can be represented in the state space form (Bosq, 2000),

$$\begin{pmatrix} Y_k \\ Y_{k-1} \\ \vdots \\ Y_{k-p+1} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \cdots & \Phi_{p-1} & \Phi_p \\ I_d & & & 0 \\ & \ddots & & \vdots \\ & & I_d & 0 \end{pmatrix} \begin{pmatrix} Y_{k-1} \\ Y_{k-2} \\ \vdots \\ Y_{k-p} \end{pmatrix} + \begin{pmatrix} \epsilon_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{2-1}$$

The operator matrix in Equation 2-1 is represented by Φ^*, and the elements I_d and 0 mean the identity operators and zero operators on H, respectively. Then Φ^* should satisfy $\|(\Phi^*)^{k_0}\|_{\mathcal{L}} < 1$ for some $k_0 \geq 1$. This condition ensures that the time series process has a strictly stationary and causal solution in L^2_H.

The prediction algorithm in Aue et al. (2015) proceeds in three steps.

Step 1. Select the number of principal components d for the observed curves. With the sample eigenfunctions, empirical fPC scores $y^e_{k,j} = \langle Y_k - \mu, \hat{v}_j \rangle$ can now be computed for each observation Y_k, $k = 1, \ldots, n$. Then we have the fPC score vectors for the kth observation

$$\mathbf{Y}^e_k = (y^e_{k,1}, \ldots, y^e_{k,d})'.$$

By nature of fPCA, the vector Y^e_k contains most of the information on the curve Y_k.

Step 2. Fix the prediction lag h. Then find a multi-dimensional time series model $\mathbf{Y}_k = \sum_{j=1}^p \Phi_j \mathbf{Y}_{k-j} + \mathbf{E}_k$ for the eigenscore vectors to produce the h-step ahead prediction

$$\hat{\mathbf{Y}}^e_{n+1} = (\hat{y}^e_{n+1,1}, \ldots, \hat{y}^e_{n+1,d})'.$$

Durbin–Levinson and innovations algorithm can be readily applied here, given the vectors $\mathbf{Y}^e_1, \ldots, \mathbf{Y}^e_n$.

Step 3. The multivariate predictions are retransformed to functional trajectories. This retransformation is achieved by defining the truncated Karhunen–Loève representation

$$\hat{Y}_{n+1} = \hat{\mu} + \hat{y}^e_{n+1,1}\hat{v}_1 + \cdots + \hat{y}^e_{n+1,d}\hat{v}_d.$$

Based on the predicted fPC scores $y^e_{k,j}$ and the estimated eigenfunctions $\hat{v}_j$, the resulting $\hat{Y}_{n+1}$ is then used as the h-step ahead functional prediction of Y_{n+1}.

2.3 Fully Functional Regression Model

In a fully functional regression model, both the explanatory "variables" (or functions) and responses are functions. Here we use multivariate technique after projection to do the estimation for the regression model. Suppose we have random explanatory functions $X(s)$ and independent response functions $Y(t)$. Denote their mean functions by $\mu_X(s) = \mathrm{E}[X(s)]$ and $\mu_Y(t) = \mathrm{E}[Y(t)]$, and their covariance functions by $C_X(s_1, s_2) = \mathrm{cov}(X(s_1), X(s_2))$, $C_Y(t_1, t_2) = \mathrm{cov}(Y(t_1), Y(t_2))$. The Karhunen–Loève expansions for the trajectories X and Y are

$$X(s) = \mu_X(s) + \sum_{i=1}^{\infty} \xi_i \phi_i(s) \qquad \text{and} \qquad Y(t) = \mu_Y(t) + \sum_{j=1}^{\infty} \zeta_j \psi_j(t),$$

where ξ_j's and ϕ_j's (ζ_j's and ψ_j's) are the fPC scores and eigenfunctions of C_X (C_Y). The fully functional linear regression model with response function Y and predictor function X can be written as

$$Y(t) = \mu_Y(t) + \int \beta(s,t)(X(s) - \mu_X(s))ds + \epsilon(t),$$

where $\epsilon(t)$'s are independent error functions, and the bivariate regression kernel $\beta(s,t)$ is assumed to be continuous and square integrable, and thus, $\int\int |\beta(s,t)|dsdt < \infty$.

The kernel function above indicates which parts of the predictor trajectory have positive vs. negative contribution to the response function $Y(t)$. Under the given assumptions, $\beta(s,t)$ has the basis representation

$$\beta(s,t) = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} \beta_{ij}\phi_i(s)\psi_j(t).$$

For simplicity we will assume the mean function of X's and Y's are both zero. Replacing $Y(t)$ and $X(s)$ with their Karhunen Loéve representation, we have

$$\sum_{j=1}^{\infty} \zeta_j\psi_j(t) = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} \beta_{ij}\xi_i\psi_j(t) + \epsilon(t).$$

For arbitrary $k \in \mathbb{Z}^+$, taking the inner product with $\psi_k(t)$ on both sides, we have

$$\zeta_j = \sum_{i=1}^{\infty} \beta_{ij}\xi_i + u_j,$$

where $u_j = \langle \epsilon, \psi_j \rangle$. In practice, we only adopt the first d_x fPCs of predictors, so we consider the following equation

$$\zeta_j = \sum_{i=1}^{d_x} \beta_{ij}\xi_i + \nu_j, \tag{2-2}$$

where $\nu_j = u_j + \sum_{i>d_x} \beta_{ij}\xi_i$. Equation 2-2 resembles a multivariate regression model. Therefore, the estimation of β_{ij} can be obtained by fitting a regression model to the d_y-dimensional eigenscore vectors of the responses against the d_x-dimensional eigenscore vectors of the explanatory functions as presented in Equation 2-2. Thus, we first estimate the eigenscores ξ's and ζ's and then estimate β_{ij}'s by fitting multiple multivariate linear regression models. From prediction perspective, we can first predict the eigenscores of Y, and obtain the predicted curve $\hat{Y}$ by truncated Karhunen-Loève expansion $\hat{Y} = \hat{\mu}_Y + \sum_{j=1}^{d_y} \hat{\zeta}_j\hat{\psi}_j$.

2.3.1 Intra-day prediction with functional regression

Without loss of generality, let Z denote a random function in $L^2[0,1]$ with mean zero. In a regression setting for intra-day prediction, the sub-curve $Z(s)|_{[0,\tau]} = (Z(s)\colon s \in [0,\tau])$ serves as the explanatory function, and the sub-curve $Z(t)|_{(\tau,1]} = (Z(t)\colon t \in (\tau,1])$ serves as response function. The Karhunen–Loève expansions of the two functional variables are

$$Z(s)|_{[0,\tau]} = \sum_{i=1}^{\infty} \xi_i^{(\tau)} \phi_i(s) \qquad \text{and} \qquad Z(t)|_{(\tau,1]} = \sum_{j=1}^{\infty} \zeta_j^{(\tau)} \psi_j(t),$$

where the notation ξ_i, ϕ_i, ζ_j and ψ_j are defined analogously to those on the entire domain $[0,1]$, but they correspond to the sub-domains $[0,\tau]$ or $(\tau,1]$.

We consider a fully functional linear regression model

$$Z(t)|_{(\tau,1]} = \int_0^{\tau} \beta_\tau(s,t) Z(s)|_{[0,\tau]} ds + \epsilon(t).$$

Here, given a fixed value of τ, assume the bivariate regression function $\beta_\tau(s,t)$ to be continuous and square integrable, consequently, $\int_\tau^1 \int_0^\tau \beta_\tau(s,t) dsdt < \infty$. By the discussion in section 2.3, the functional regression model can be expressed as

$$Z(t)|_{(\tau,1]} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \beta_{\tau,ij} \xi_i^{(\tau)} \psi_j(t) + \epsilon(t),$$

where $\beta_{\tau,ij}$ are the regression parameters to be estimated. Under the continuity assumption on $\beta_\tau(s,t)$ along with τ, it follows that $\beta_{\tau,ij}$ is also continuous in τ for all i and j. In the following section, we will introduce a novel criterion that allow us to jointly select the number of principal components for predictors and responses.

2.3.2 Dimension selection for fully functional regression model

Typically, we will project the functional objects into a finite dimensional space spanned by the first few principal components. The number of principal components are selected such that the proportion of variance explained exceeds a prespecified threshold (say, 90%). However, our purpose is prediction, so it is not always appropriate to select principal components that explain a large portion of variance. So we consider a new criterion for selecting the best dimensions of eigenfunction spaces of predictors and responses. Here, we propose to choose the dimensions by minimizing the mean square error of prediction.

Without loss of generality, assume predictors X's and responses Y's be elements in L_H^2 with mean function zero and covariance operator C_X resp. C_Y. Suppose the dimension of eigenfunction space of the predictors is d_x, that of the responses is d_y, and $\hat{Y}$ is the prediction of Y by the regression model, then by orthonormality of eigenfunctions, the MSE of prediction can be decomposed as

$$\mathrm{E}[\|Y_{n+1} - \hat{Y}_{n+1}\|^2] = \mathrm{E}[\|\mathbf{Y}_{n+1} - \hat{\mathbf{Y}}_{n+1}\|^2] + \sum_{l>d_y} \lambda_l^Y,$$

where $\mathbf{Y}_k = (\zeta_{k1}, \ldots, \zeta_{kd_y})'$ is the truncated eigenscore vector of the curve to be predicted, $\hat{\mathbf{Y}}_k = (\hat{\zeta}_{k1}, \ldots, \hat{\zeta}_{kd_y})'$ is the prediction of $\mathbf{Y}$, and λ_l^Y is the lth eigenvalue of C_Y, and $\|\cdot\|$ denotes the Euclidean norm.

Let $\mathbf{X}_k = (\xi_{k1}, \ldots, \xi_{kd_x})'$ be the truncated eigenscore vector of the predictors, then by the discussion above, there exists a $d_y \times d_x$ matrix $B = \{\beta_{ij}\}_{i,j=1}^{d_x,d_y}$, such that $\mathbf{Y}_k = B\mathbf{X}_k + \mathbf{Z}_k$, where $\mathbf{Z}_k = (z_{k1}, \ldots, z_{kd_y})'$ with $z_{kj} = \sum_{i>d_x} \beta_{ij}\xi_{ki} + \langle \epsilon_k, \psi_j \rangle$, where ψ_j is the

jth eigen-function of C_Y. We assume the covariance matrix of $\mathbf{Z}_k$ to be Σ_z. Therefore,

$$\begin{aligned} \mathrm{E}[\|\mathbf{Y}_{n+1} - \hat{\mathbf{Y}}_{n+1}\|^2] &= \mathrm{E}[\|\mathbf{Y}_{n+1} - \hat{B}\mathbf{X}_{n+1}\|^2] \quad (2\text{-}3) \\ &= \mathrm{E}[\|(B - \hat{B})\mathbf{X}_{n+1}\|^2] + \mathrm{E}[\|\mathbf{Z}_{n+1}\|^2] \quad (2\text{-}4) \\ &= \mathrm{E}[\|(\mathbf{X}'_{n+1} \otimes I_{d_y})(\beta - \hat{\beta})\|^2] + \mathrm{E}[\|\mathbf{Z}_{n+1}\|^2]. \quad (2\text{-}5) \end{aligned}$$

Let $\tilde{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, $\tilde{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, $\tilde{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, $\beta = \mathrm{vec}(B)$, and $\hat{\beta} = \mathrm{vec}(\hat{B})$ be its least squares estimator. Then we have $\tilde{Y} = B\tilde{X} + \tilde{Z}$, or equivalently,

$$\tilde{y} = (\tilde{X}' \otimes I_{d_y})\beta + \tilde{z},$$

where $\tilde{y} = \mathrm{vec}(\tilde{Y})$ and $\tilde{z} = \mathrm{vec}(\tilde{Z})$. The least squares estimator of β is

$$\hat{\beta} = ((\tilde{X}\tilde{X}')^{-1}\tilde{X} \otimes I_{d_y})\tilde{y},$$

and we also have

$$\hat{\beta} - \beta = ((\tilde{X}\tilde{X}')^{-1}\tilde{X} \otimes I_{d_y})\tilde{z}.$$

Our next task is to study the asymptotic property of $\hat{\beta}$. Following the above equation,

$$\begin{aligned} \sqrt{N}(\hat{\beta} - \beta) &= \sqrt{N}((\tilde{X}\tilde{X}')^{-1}\tilde{X} \otimes I_{d_y})\tilde{z} \\ &= \left(\left(\frac{1}{N}\tilde{X}\tilde{X}'\right)^{-1} \otimes I_{d_y}\right) \frac{1}{\sqrt{N}} \left(\tilde{X} \otimes I_{d_y}\right) \tilde{z}. \end{aligned}$$

By the weak law of large number, we have $\left(\frac{1}{N}\tilde{X}\tilde{X}'\right)^{-1} \otimes I_{d_y} \xrightarrow{p} \Sigma_x^{-1} \otimes I_{d_y}$, where Σ_x is the covariance matrix of X_n's. By the central limit theorem,

$$\begin{aligned} \frac{1}{\sqrt{N}} \left(\tilde{X} \otimes I_{d_y}\right) \tilde{z} &= \frac{1}{\sqrt{N}}\mathrm{vec}(\tilde{Z}\tilde{X}') \\ &\xrightarrow{d} \mathcal{N}\left(0, \Sigma_x \otimes \Sigma_z\right). \end{aligned}$$

Finally, by Slutsky's theorem,

$$\sqrt{N}\left(\hat{\beta}-\beta\right)\overset{d}{\to}\mathcal{N}\left(0,\Sigma_x^{-1}\otimes\Sigma_z\right). \tag{2-6}$$

As for the first term in Equation 2-5, it is reasonable to assume that $\hat{\beta}$ and $\mathbf{X}_{n+1}$ are independent since asymptotically the sample size will go to infinity, and $\hat{\beta}$ is based on the whole sample, so the dependence between $\hat{\beta}$ and $\mathbf{X}_{n+1}$ is negligible for large sample sizes. Then by the independence and Equation 2-6,

$$\begin{aligned}
\mathrm{E}[\|(\mathbf{X}'_{n+1}\otimes I_{d_y})(\beta-\hat{\beta})\|^2] &= \mathrm{tr}\{\mathrm{E}[(\mathbf{X}_{n+1}\mathbf{X}'_{n+1}\otimes I_{d_y})(\beta-\hat{\beta})(\beta-\hat{\beta})']\}\\
&= \mathrm{tr}\{(\Sigma_x\otimes I_{d_y})\mathrm{E}[(\beta-\hat{\beta})(\beta-\hat{\beta})']\}\\
&= \frac{1}{n}\mathrm{tr}\{(\Sigma_x\otimes I_{d_y})(\Sigma_x^{-1}\otimes\Sigma_z)+o(1)\}\\
&\sim \frac{1}{n}\mathrm{tr}(I_{d_x}\otimes\Sigma_z)\\
&= \frac{d_x}{n}\mathrm{tr}(\Sigma_z),
\end{aligned}$$

where $a_n \sim b_n$ means $a_n/b_n \to 1$. It can be shown $\mathrm{E}[\|\mathbf{Z}_{n+1}\|^2] = \mathrm{tr}(\Sigma_z)$. Therefore, we have

$$\mathrm{E}[\|Y_{n+1}-\hat{Y}_{n+1}\|^2] \sim \frac{n+d_x}{n}\mathrm{tr}(\Sigma_z)+\sum_{l>d_y}\lambda_l^Y.$$

Replacing λ_l^Y with $\hat{\lambda}_l^Y$, and $\mathrm{tr}(\Sigma_z)$ with $\frac{n}{n-d_x}\mathrm{tr}(\hat{\Sigma}_z)$ as $\mathrm{E}\left[\frac{1}{n-d_x}\hat{Z}\hat{Z}'\right]=\Sigma_z$, we have the fFPE criterion for fully functional regression model shown as follows:

$$\mathrm{fFPE_r}(d_x,d_y)=\frac{n+d_x}{n-d_x}\mathrm{tr}(\hat{\Sigma}_z)+\sum_{l>d_y}\hat{\lambda}_l^Y.$$

Then it is natural to propose to choose d_x and d_y by minimizing the above objective function. The following theorem shows the consistency of the criterion.

Theorem 1. Suppose $(X_k\colon k\in\mathbb{N})\in L^2[a,b]$, $(Y_k\colon k\in\mathbb{N})\in L^2[c,d]$ are two series of

L^4-m approximable (see Hörmann and Kokoszka (2010)) random functions satisfying $E[\|X_k\|^{4+\epsilon}] < \infty$ and $E[\|Y_k\|^{4+\epsilon}] < \infty$ for some $\epsilon > 0$, serving as predictor and responses in a fully functional regression model

$$Y_k(t) = \int \beta(t,s)X_k(s)ds + \epsilon_k(t),$$

and $\hat{Y}_{n+1}$ is the prediction of Y_{n+1} based on C_X and C_Y, and $\tilde{Y}_{n+1}$ is the prediction of Y_{n+1} based on $\hat{C}_X$ and $\hat{C}_Y$ and $\hat{c}_j = \text{sign}\langle \phi_j, \hat{\phi}_j \rangle$, $\hat{d}_j = \text{sign}\langle \psi_j, \hat{\psi}_j \rangle$. Then if $E[Y^4(t) \otimes Y^4(s)] < \infty$ for arbitrary t, we have

$$E[\|Y_{n+1} - \hat{Y}_{n+1}\|^2] - E[\|Y_{n+1} - \tilde{Y}_{n+1}\|^2] \to 0, \qquad \text{as } n \to \infty.$$

3 Proposed Prediction Method

We know the following decomposition framework for smooth trajectories,

$$Y_k(t) = S_k(t) + \epsilon_k(t), \qquad t \in [0,1],$$

where $S(t)$ is the signal correlated to the previous curves, and $\epsilon(t)$ is the innovation function independent with the previous curves. Further, if the observed curves are contaminated by random noise, we can decompose the functional time series into three parts:

$$Y_k(t_j) = S_k(t_j) + \epsilon_k(t_j) + e_k(t_j), \qquad k = 1, \ldots, n, \; j = 1, \ldots, J,$$

where $e(t_j)$ represents random error. In practice, the observations are available only at prespecified discrete grids, so here we use t_j instead of t. We propose a stage-wise procedure, where each stage corresponds to predicting one component and combine them to obtain the final prediction.

3.1 Smooth functions

For any function $Y_k(t)$, the trajectory over $[0,\tau]$ is denoted by $Y_k|_{[0,\tau]}$, and the trajectory over $(\tau,1]$ is denoted by $Y_k|_{(\tau,1]}$. Suppose we have observed $Y_1,\dots,Y_n$, and $Y_{n+1}|_{[0,\tau]}$. The updated prediction of the curve over $(\tau,1]$ is given by

$$\hat{Y}^u_{n+1}|_{(\tau,1]} = \hat{Y}_{n+1}|_{(\tau,1]} + \hat{\epsilon}_{n+1}|_{(\tau,1]},$$

where $\hat{Y}_{n+1}$ is the "next-interval" prediction of Y_{n+1} and $\hat{\epsilon}_{n+1}|_{(\tau,1]}$ the intraday prediction of the $(n+1)$th innovation function over sub-domain $(\tau,1]$.

To predict $\epsilon_{n+1}|_{(\tau,1]}$, we consider a fully functional regression model, where $(\epsilon_i(s)|_{[0,\tau]})_{i=1}^n$ serve as the "predictors" and $(\epsilon_i(t)|_{(\tau,1]})_{i=1}^n$ serve as the responses,

$$\epsilon_k(t)|_{(\tau,1]} = \int_0^\tau \beta_\tau(s,t)\epsilon_k(s)|_{[0,\tau]}ds + \delta_k(t).$$

By the Karhunen–Loève expansion,

$$\epsilon_k(s)|_{[0,\tau]} = \sum_{i=1}^\infty \xi_i^{(k)}\phi_i(s)|_{[0,\tau]} \qquad \text{and} \qquad \epsilon_k(t)|_{(\tau,1]} = \sum_{j=1}^\infty \zeta_j^{(k)}\psi_j(t)|_{(\tau,1]}.$$

The innovation function is unobserved, so we apply the functional regression model to the residual in the first step $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i$ is the full-curve prediction. Replacing the unknown terms with the estimated values, and adopting the first d_x and d_y fPCs for predictors and responses respectively, we have

$$\hat{\hat{\epsilon}}_{n+1}(t)|_{(\tau,1]} = \sum_i^{d_x}\sum_j^{d_y} \hat{\beta}_{\tau,ij}\hat{\xi}_i^{(n+1)}\hat{\psi}_j(t)|_{(\tau,1]},$$

where $\hat{\xi}_i^{(n+1)} = \langle \hat{\epsilon}_{n+1}|_{[0,\tau]}, \hat{\phi}_i\rangle$. Then $\hat{\hat{\epsilon}}_{n+1}(t)|_{(\tau,1]}$ is the prediction of $\epsilon_{n+1}(t)|_{(\tau,1]}$ given

$\hat{\epsilon}_{n+1}(s)|_{[0,\tau]}$. Therefore the final prediction of $Y_{n+1}|_{(\tau,1]}$ is

$$\hat{Y}^u_{n+1}|_{(\tau,1]} = \hat{Y}_{n+1}|_{(\tau,1]} + \hat{\hat{\epsilon}}_{n+1}|_{(\tau,1]}.$$

The updated prediction $\hat{Y}^u_{n+1}|_{(\tau,1]}$ can be regarded as the complete curve prediction $\hat{Y}_{n+1}|_{(\tau,1]}$ adjusted by the intra-day prediction of the $(\tau, 1]$ block of the residual function $\hat{\hat{\epsilon}}_{n+1}|_{(\tau,1]}$. The prediction steps can be summarized by the following algorithm.

Step 1. Fix d, p, and apply functional time series prediction (e.g. Aue et al. (2015)), to obtain the prediction $\hat{Y}_{n+1}$ for Y_{n+1}.

Step 2. Obtain the prediction residual functions $\hat{\epsilon}_k$'s for a training group $\{Y_k\}^n_{k=n_1}$, where the window size for the prediction of each curve in the training group is n_1.

Step 3. Separate the prediction residual functions in Step 2 at "current time" τ. Treat the first parts $(\hat{\epsilon}_k|_{[0,\tau]})^n_{k=n_1}$ as the predictors, and the second parts $(\hat{\epsilon}_k|_{(\tau,1]})^n_{k=n_1}$ as the responses. Fix d_x and d_y, and apply intra-day functional regression model on the second segments $(\hat{\epsilon}_k|_{(\tau,1]})^n_{k=n_1}$ against the first segments $(\hat{\epsilon}_k|_{[0,\tau]})^n_{k=n_1}$, and use the fitted model to obtain the prediction of the $(\tau, 1]$ block of the Y_{n+1}'s residual function $\hat{\hat{\epsilon}}_{n+1}|_{(\tau,1]}$.

Step 4. Add the $(\tau, 1]$ segment of the complete predicted curve $\hat{Y}_{n+1}$ and the predicted $(\tau, 1]$ block of the residual function to get the final prediction $\hat{Y}^u_{n+1}|_{(\tau,1]} = \hat{Y}_{n+1}|_{(\tau,1]} + \hat{\hat{\epsilon}}_{n+1}|_{(\tau,1]}$.

3.2 Noisy functions

In this section, we consider functional data as noisy sampled points from a collection of consecutive trajectories. In practice, the observed functional time series is observed at a discrete time grid, thus the observed curves can be rough. The reasons may be measurement errors or sparsely-spaced observation time grids. As has been discussed by Yao et al. (2005), the rough error term will lead to biased fPC scores, so we need to prevent the problem. In practice, we can use some smooth basis functions to smooth the raw trajectories. However, in the random error $(e_k(t_j), k \in \mathbb{Z}, j \in 1, \dots, l)$, which is not smooth, there could still exist short-term time series correlation, so we need one more step to extract the information left in the pre-smoothing residuals. Because the time dependency in the random error usually decays very fast as lag increases, we can only expect reasonable predictions for the near future.

As has been discussed, we decompose any functional time series $(Y_i(t), i \in \mathbb{Z})$ into three parts,

$$Y_k(t_j) = S_k(t_j) + \epsilon_k(t_j) + e_k(t_j), \qquad k \in \mathbb{Z},\ j = 1, \dots, J,$$

where $S_k(t_j)$ is the smooth signal from the smooth part of the past time series observations, $\epsilon_k(t_j)$ is the independent smooth innovation function, and $e_k(t_j)$ is the random error of the functional time series.

Let $f_k(t_j) = S_k(t_j) + \epsilon_k(t_j)$ represent the smooth part of the functional time series, which can be predicted by functional methodology, while $e_k(t_j)$ is the rough part. If there is time series correlation in this process, it can be predicted by ARMA model. Here we apply ARMA model to the pre-smoothing residual $\{r_k(t_j)\}$, defined as $r_k(t_j) = \widetilde{Y}_k(t_j) - Y_k(t_j)$, where $\widetilde{Y}_k(t_j)$ is the original time series, and $Y_k(t)$ is the smoothed functional time series. For noisy trajectories, we add two more steps to the previous algorithm.

Step 5. Apply ARMA model to the pre-smoothing residuals, to predict the future residuals $\hat{r}_{n+1}(t_j)$.

Step 6 Combine the prediction of the smooth part in Step 4 and pre-smoothing residuals to obtain the final prediction.

$$\hat{Y}_{n+1}(t_j) = \hat{f}_{n+1}(t_j) + \hat{r}_{n+1}(t_j).$$

The final prediction for Y_{n+1} is $\hat{Y}_{n+1}(t_j) = \hat{f}_{n+1}(t_j)+\hat{r}_{n+1}(t_j)$, where $t_j = (1/l, 2/l, \ldots, (l-1)/l, 1)$, $\hat{r}_{n+1}(t_j)$ is ARMA prediction of $\{r_{n+1}(t_j)\}$. This adjustment is necessary if the observed functional time series curves are significantly rough and time series structure in $r_k(t_j)$'s is pronounced. The prediction of the smooth part can be also viewed as a de-trending process. As has been shown in the appendix, the autocorrelation of the pre-smoothing residuals decays much faster than that of the original time series, which indicates that the long-term dynamics (e.g. seasonal trend) has been removed.

3.3 Selection of p, d, d_x and d_y

We develop a method for the selection of unknown parameters that is based on the prediction error, the order and the dimension of the projected eigenspace at the first stage will influence the covariance function of the residual functions, which will further influence the intra-day prediction. Therefore, $\hat{\Sigma}_\delta$ and $\hat{\lambda}_j^{\epsilon_{T(\tau)}}$ can be regarded as functions of p and d and thus we propose to jointly select p, d, d_x and d_y by minimizing the following objective function

$$\text{fFPE}(p, d, d_x, d_y) = \frac{n + d_x}{n - d_x}\text{tr}(\hat{\Sigma}_\delta(p, d)) + \sum_{l>d_y} \hat{\lambda}_j^{\epsilon_{T(\tau)}}(p, d),$$

With the use of this functional FPE criterion, PFP is fully data-driven and we do not need additional tuning parameter adjustment.

4 Simulation

4.1 General setting

To analyze the finite sample properties of the new prediction method, a comparative simulation study was conducted. PFP was tested on simulated FAR models. In each simulation test, 400 curves were generated. Beginning from the first curve, the following consecutive 200 trajectories were used as the training group to obtain the residual function of the one-step ahead prediction. Then we switched the training group with the same number of functions in a sliding window way, to obtain the prediction residual function for the next curve. Finally we had 200 estimated prediction residual functions, among which the first 180 functions were fitted by an intraday functional regression model, which was used to predict the unobserved block of the rest 20 curves. The corresponding mean square error of prediction was computed, as well as the fFPE value for comparison. This procedure was repeated for 100 times for each simulation run.

In the simulation, we worked in a D-dimensional functional space H, which is spanned by D Fourier basis functions $\mathbf{v} = (\nu_1, \nu_2, \ldots, \nu_D)$ on the unit interval $[0, 1]$. Any arbitrary element in H has the representation $x(t) = \sum_{j=1}^{D} c_j \nu_j(t)$ with coefficients $\mathbf{c} = (c_1, \ldots, c_D)'$. Then for any linear operator $\Psi \colon H \to H$, we have

$$\Psi(x) = \sum_{j=1}^{D} c_j \Psi(\nu_j) = \sum_{j=1}^{D} \sum_{j'=1}^{D} c_j \langle \Psi(\nu_j), \nu_{j'} \rangle \nu_{j'} = \mathbf{c}' \boldsymbol{\Psi} \mathbf{v},$$

where $\boldsymbol{\Psi}$ is a $D \times D$ matrix with elements $(\langle \Psi(\nu_j), \nu_{j'} \rangle)_{j,j'=1}^{D}$. The linear operator used to generate FAR model then can be represented in matrix form. The innovation function is generated by $\epsilon_k(t) = \sum_{j=1}^{D} a_{k,j} c_j$, where $a_{k,j}$'s are i.i.d. normal random variables with mean zero and standard deviation σ_j. Two sets of standard deviations used here are $\sigma_1 = (j^{-1} \colon j = 1, \ldots, D)$ and $\sigma_2 = (1.2^{-j} \colon j = 1, \ldots, D)$.

4.2 Prediction comparison for smooth curves

In this section, we show the comparison of partial functional prediction with Aue et al. (2015)'s method and intraday functional regression method on FAR(2) processes $Y_k = \Psi_1 Y_{k-1} + \Psi_2 Y_{k-2} + \epsilon_k$. We assume the $(\tau, 1]$ part of the last 20 trajectories is unobserved and the $[0, \tau]$ part is observed, so we only need to predict the unobserved part of these curves.

The operators were generated such that $\Psi_1 = \kappa_1 \Psi$ and $\Psi_2 = \kappa_2 \Psi$. (Here, note that $\kappa_2 = 0$ yields a FAR(1) process). The operator matrix $\mathbf{\Psi}$ is generated at random, with each element following a normal distribution with mean zero and variance $\sigma_{ll'}$, and then scaled by its l_2 norm. In each simulation run, the operator matrix is newly generated. We chose $\sigma_{jj'}$ to be $(\sigma_i \sigma_i')_{jj'}$ to ensure the simulated functions satisfying Riemann-Lebesgue Lemma. We set $D = 15$ in our simulation.

In each simulation run, the MSE of prediction

$$\int_{\tau}^{1} [Y_{n+1}(t) - \hat{Y}_{n+1}(t)|_{(\tau,1]}]^2 dt$$

of PFP and the three competitor method: time series method (Aue et al. (2015)), functional mixture method (Chiou 2012), and intraday functional regression are computed. The fFPE values are also calculated for the partial functional prediction and the intraday regression method, which is recorded to be close to the corresponding MSE of prediction. For the competitor time series method, we do not provide the fFPE value since Aue et al. (2015) have shown they should be close to the MSE of prediction. Results for five pairs of values (κ_1, κ_2) are provided in Table 1.

We find that when time series structure is strong, PFP will outperform the other methods. When time series structure is weak, the performances of PFP and functional linear regression (FLR) method are similar, and are better than full curve prediction

		σ_1					
κ_1	κ_2	fFPE$_{PFP}$	PMSE$_{PFP}$	PMSE$_{ts}$	PMSE$_i$	fFPE$_r$	PMSE$_r$
1.8	0.0	0.2024	0.2097	0.8442	0.6428	0.3269	0.3431
0.8	0.0	0.2003	0.2112	0.8396	0.5779	0.2664	0.2763
0.2	0.0	0.1928	0.2018	0.8286	0.4622	0.1938	0.2025
0.4	0.4	0.2038	0.2123	0.8388	0.5249	0.2309	0.2392
0.0	0.8	0.2058	0.2115	0.8419	0.5977	0.2647	0.2685
		σ_2					
κ_1	κ_2	fFPE$_{PFP}$	PMSE$_{PFP}$	PMSE$_{ts}$	PMSE$_i$	fFPE$_r$	PMSE$_r$
1.8	0.0	0.5554	0.5801	1.2269	1.9770	1.1012	1.1668
0.8	0.0	0.5455	0.5640	1.2112	1.1966	0.7011	0.7431
0.2	0.0	0.5302	0.5561	1.1813	1.0035	0.5287	0.5536
0.4	0.4	0.5711	0.5985	1.2593	1.0634	0.6128	0.6391
0.0	0.8	0.5740	0.5907	1.2631	1.2516	0.6995	0.7127

Table 1: Average fFPE values and prediction MSEs for different pairs of κ_1 and κ_2 from 100 iterations of the three methods, fFPE$_{PFP}$ and PMSE$_{PFP}$ are the fFPE value and prediction MSE of PFP, fFPE$_r$ and PMSE$_r$ are the fFPE value and prediction MSE of intra-day regression method, and PMSE$_{ts}$ is the prediction MSE of time series prediction method. PMSE$_i$ is the prediction MSE of Chiou's functional mixture prediction method, and the number of clusters is 3, and we set $\tau = 0.5$ in each case.

method and functional mixture method. The fFPE value and the prediction MSE are always very close for different situations. This numerically approves the practical applicability of the fFPE criterion.

4.2.1 Empirical validity of the fFPE criterion

In this section, p, d, d_x, and d_y are selected jointly by the new fFPE criterion. We simulated 100 times for each setting and then took the average of fFPE value and prediction MSE for comparison. In Table 2, we show the selected order and dimensions and the minimal fFPE value (denoted by fFPE$_a$), the minimal prediction MSE (denoted by PMSE$_b$), the fFPE value corresponding to the minimal prediction MSE (denoted by fFPE$_b$), and the prediction MSE corresponding to minimal fFPE value (denoted by PMSE$_a$).

It is clear that the fFPE$_a$ and fFPE$_b$ values are very close, and that PMSE$_a$ and PMSE$_b$

		σ_1							
κ_1	κ_2	p	d	d_x	d_y	fFPE$_a$	fFPE$_b$	PMSE$_a$	PMSE$_b$
1.8	0.0	1	7	12	12	0.1964	0.1964	0.2021	0.2021
0.8	0.0	1	5	12	12	0.1943	0.1944	0.2016	0.2011
0.2	0.0	1	5	12	12	0.1901	0.1901	0.1979	0.1979
0.4	0.4	2	5	12	12	0.1964	0.1964	0.2036	0.2036
0.0	0.8	2	5	12	12	0.1980	0.1980	0.2035	0.2035
		σ_2							
κ_1	κ_2	p	d	d_x	d_y	fFPE$_a$	fFPE$_b$	PMSE$_a$	PMSE$_b$
1.8	0.0	1	10	12	12	0.5520	0.5520	0.5754	0.5754
0.8	0.0	1	8	12	12	0.5429	0.5429	0.5599	0.5599
0.2	0.0	1	2	12	12	0.5279	0.5296	0.5553	0.5526
0.4	0.4	2	6	12	12	0.5636	0.5636	0.5759	0.5759
0.0	0.8	2	8	12	12	0.5573	0.5575	0.5783	0.5768

Table 2: Selected order and dimensions for different choices of κ_1 and κ_2 and the average fFPE and prediction MSE from 100 iterations. We set $\tau = 0.5$ for each case.

are also very close. This confirms that in practice it will be sensible to jointly select the dimensions and order by this fFPE criterion. Even though the PMSE does not necessarily reach its minimum with the same pair of p, d with which fFPE value reaches its minimum, the minimal PMSE and the PMSE corresponding to the minimal fFPE value are still very close. Thus, the fFPE criterion may not always give the best order and dimensions, but can avoid bad selection, and the selected parameters should be close to the best ones.

4.3 Prediction comparison for noisy curves

We simulates a series of rough functional time series by adding AR(1) errors to the smooth functional time series. We set $\kappa_1 = 1.8$ and $\kappa_2 = 0$. Then the simulated functions are

$$Y_k(t_j) = S_k(t_j) + e_k(t_j),\ j = 1, \ldots, 48$$

where $S_k(t_j)$ is the smoothed curve obtained from the simulated FAR(1) process and $e_k(t_j)$ is the AR(1) error. The "current time" $\tau = 0.5$. The average prediction error of the following 5 grids ($1 \leq h \leq 5$) of the last 20 curves are shown in Table 3.

	$\phi = 0.5$, $\sigma = 0.2$				$\phi = 0.5$, $\sigma = 0.5$				$\phi = 0.8$, $\sigma = 0.5$			
h	MSEc	MSEn	MSEa	MSEi	MSEc	MSEn	MSEa	MSEi	MSEc	MSEn	MSEa	MSEi
h=1	0.2692	0.8986	0.3965	0.2950	0.4917	0.9362	0.6295	0.6665	0.4657	0.9661	0.5919	0.6468
	(0.359)	(0.083)	(0.231)	(0.327)	(0.325)	(0.190)	(0.245)	(0.240)	(0.350)	(0.150)	(0.260)	(0.240)
h=2	0.3711	0.6648	0.8446	0.4572	0.5365	0.6777	0.9872	0.8138	0.5307	0.7050	0.9745	0.8612
	(0.392)	(0.110)	(0.159)	(0.339)	(0.355)	(0.225)	(0.175)	(0.245)	(0.425)	(0.170)	(0.170)	(0.235)
h=3	0.3297	0.3159	1.3165	0.6106	0.4715	0.4275	1.3991	0.9037	0.4819	0.4280	1.4347	1.0113
	(0.338)	(0.363)	(0.080)	(0.219)	(0.320)	(0.395)	(0.110)	(0.175)	(0.370)	(0.370)	(0.095)	(0.165)
h=4	0.2189	0.4064	1.8303	0.7634	0.4083	0.4678	1.8587	0.9607	0.3645	0.4560	1.8708	1.0901
	(0.565)	(0.246)	(0.043)	(0.146)	(0.450)	(0.310)	(0.060)	(0.180)	(0.455)	(0.340)	(0.055)	(0.150)
h=5	0.1560	0.7263	2.4103	0.9459	0.4065	0.7583	2.3927	1.1331	0.3886	0.7587	2.4336	1.2443
	(0.733)	(0.130)	(0.037)	(0.100)	(0.540)	(0.210)	(0.060)	(0.190)	(0.545)	(0.235)	(0.060)	(0.160)

Table 3: Average prediction MSE of the fPFPs, and the proportions of cases where the corresponding prediction MSE is minimum are shown in the parenthesis. MSEc is the prediction MSE of PFP in the noisy case, MSEn is the prediction MSE of PFP in the smooth case, MSEa is the prediction MSE of ARIMA model, MSEi is the prediction error of PFP with the selected pre-smoothing method being linear interpolation. We set $\tau = 0.5$ for each case. The parameter ϕ is the coefficient of the AR process of the error time series and σ^2 is the variance of the error. The simulated prediction process was repeated 200 times.

The simulation experiments indicate that the ARMA model should be the "last-resort" method to use for long-term prediction. Since the ARMA model may provide reasonable short term prediction, one can use this approach it to predict the rough errors. However, if we incorporate the error term into PFP in the smooth case by linear interpolation, the prediction will deteriorate since the estimation of the actual fPC scores is biased and this error propagates to the estimated FAR model.

4.4 Nonparametric bootstrap prediction interval

Prediction intervals are useful in practice for assessing the prediction uncertainty and accuracy. To provide the prediction interval for $Y_{n+1}|_{(\tau,1]}$, we used the same bootstrap resampling method in Chiou (2012) to the estimated innovation function which we briefly describe. Suppose that each prediction residual function has the Karhunen–Loéve representation $\hat{e}(t) = \hat{\mu}_e + \sum_{j=1}^{\infty} \hat{\xi}_j \hat{\phi}_j(t)$, and obtain $\hat{\epsilon}_i^e(t) = \hat{e}_i(t) - \hat{\mu}_e - \sum_{j=1}^{d_e} \hat{\xi}_{ij}\hat{\phi}_j(t),\ 1 \leq i \leq n$. The optimal number d_e is selected to be the smallest number such that the variance explained by the first d_e principal components exceeds a prespecified threshold (say, 80%). Then the bootstrap sample of the fPC scores $\{\hat{\boldsymbol{\xi}}_1^b, \ldots, \hat{\boldsymbol{\xi}}_n^b\}$ and the residuals $\hat{\epsilon}_1^b(t), \ldots, \hat{\epsilon}_n^b(t)$ are obtained by sampling with replacement from $\{\hat{\boldsymbol{\xi}}_i, 1 \leq i \leq n\}$ and $\{\hat{\epsilon}_i^e, 1 \leq i \leq n\}$, respectively, where $\hat{\boldsymbol{\xi}}_i = \{\hat{\xi}_{i1}, \ldots, \hat{\xi}_{id_e}\}$. The B bootstrap samples for innovations $\{\hat{e}_1^b(t), \ldots, \hat{e}_n^b(t),\ 1 \leq b \leq B\}$ are the summation

$$\hat{e}_i^b(t) = \sum_{j=1}^{d} \hat{\xi}_{ij}^b \hat{\phi}_j(t) + \hat{\epsilon}_i^b(t),\ 1 \leq i \leq n,\ 1 \leq b \leq B.$$

The final bootstrap prediction is $\hat{Y}_{n+1}^{u,b}(t)|_{(\tau,1]} = \hat{Y}_{n+1}(t)|_{(\tau,1]} + \hat{\hat{e}}_{n+1}^b(t)|_{(\tau,1]}$, where

$$\hat{\hat{e}}_{n+1}^b(t)|_{(\tau,1]} = \int_0^{\tau} \hat{\beta}^b(t,s)\hat{e}_{n+1}(s)|_{[0.\tau]} ds,\ b = 1, \ldots, B,$$

and $\hat{\beta}^b(t,s)$ is the estimated coefficient kernel function of $\beta(t,s)$ from bootstrap samples. For the B bootstrap samples, the $100(1-\alpha)\%$ pointwise prediction bands are defined as $\alpha/2 \times 100$ and $(1-\alpha/2) \times 100$ empirical pointwise percentiles of $\{\tilde{Y}_{n+1}^1(t), \ldots, \tilde{Y}_{n+1}^B(t), t \in T(\tau)\}$,

$$P\left(\hat{\xi}_l(\alpha,t) < Y(t) < \hat{\xi}_u(\alpha,t), \text{for all } t \in [0,1]\right) \approx \alpha.$$

To evaluate the interval forecast accuracy, we utilized the interval score proposed in Gneiting & Raftery (2007), given as follows

$$\begin{aligned} & S_\alpha(u(t), l(t), Y_n(t)) \\ & \quad = (u(t) - l(t)) + \frac{2}{\alpha}(Y_n(t) - u(t))\mathbf{1}\{Y_n(t) > u(t)\} + \frac{2}{\alpha}(l(t) - Y_n(t))\mathbf{1}\{\ell(t) > Y_n(t)\}, \end{aligned}$$

where $u(t)$ is the upper bound, and $\ell(t)$ is the lower bound of the prediction interval of $Y_n(t)$.

All the curves were evaluated at 48 equally-spaced grids, and we assumed the trajectory to be predicted was observed over the partial interval $[0, \tau) \subset [0, 1]$, and we set $\alpha = 0.05$. Then we obtained the bootstrap prediction interval for the 20 predicted curves of PFP and intraday prediction respectively and then averaged the scores over all grids and days to obtain the averaged score defined by

$$\bar{S}_\alpha = \frac{1}{\#\{t_j \in [\tau, 1]\} \times 20} \sum_{\#\{t_j \in [\tau,1]\}} \sum_{k=1}^{20} S_\alpha(u_k(t_j), l_k(t_j), Y_{n+k,T(\tau)}(t_j)).$$

The results are shown in the Table 4 (FLR represents functional linear regression), and the average width of the prediction interval is shown in Table 5,

$$\frac{1}{\#\{t_j \in [\tau, 1]\} \times 20} \sum_{t_j \in [\tau,1]} \sum_{k=1}^{20} \left(\hat{\xi}_{u,k}(\alpha, t_j) - \hat{\xi}_{l,k}(\alpha, t_j) \right).$$

These results show that the bootstrap prediction bands of PFP is narrower than that of functional regression model. After removing the time series dependency in the data, the variation in the predicted curve was reduced which demonstrates another advantage of PFP. The prediction bands are also provided in analysis of the pollution data and traffic data in Section 5.

σ_1		τ=0.375		τ=0.5		τ=0.625	
κ_1	κ_2	score$_{PFP}$	score$_{FLR}$	score$_{PFP}$	score$_{FLR}$	score$_{PFP}$	score$_{FLR}$
1.8	0.0	13.6638	16.1626	7.9216	11.9808	11.7451	11.2181
0.8	0.0	13.7603	16.9235	7.2512	10.0672	11.1408	11.2278
0.2	0.0	13.2666	14.1214	7.1535	8.6642	11.0978	11.9593
0.4	0.4	13.8709	14.3918	7.7111	8.6484	11.6376	11.9699
0.0	0.8	13.8105	14.5196	7.7204	8.1962	12.3002	11.9588
σ_2		τ=0.375		τ=0.5		τ=0.625	
κ_1	κ_2	score$_{PFP}$	score$_{FLR}$	score$_{PFP}$	score$_{FLR}$	score$_{PFP}$	score$_{FLR}$
1.8	0.0	21.7463	27.0548	13.4178	23.9143	19.8813	19.5835
0.8	0.0	22.1516	24.6587	14.5521	17.7493	21.1014	20.8474
0.2	0.0	21.4360	21.6423	14.5120	15.2796	21.1205	21.2764
0.4	0.4	21.9319	21.8376	14.6626	16.0734	21.4654	21.0682
0.0	0.8	21.3932	21.3133	14.7013	16.1025	21.6377	20.1374

Table 4: Interval scores for different choices of κ_1 and κ_2 from 1000 bootstrap iterations

σ_1		τ=0.375		τ=0.5		τ=0.625	
κ_1	κ_2	score$_{PFP}$	score$_{FLR}$	score$_{PFP}$	score$_{FLR}$	score$_{PFP}$	score$_{FLR}$
1.8	0.0	0.5318	0.8678	0.4579	0.7581	0.5036	0.6035
0.8	0.0	0.5366	0.6061	0.4491	0.5400	0.4999	0.5002
0.2	0.0	0.5163	0.5120	0.4406	0.4390	0.4859	0.4794
0.4	0.4	0.5068	0.5167	0.4399	0.4589	0.4840	0.4756
0.0	0.8	0.5132	0.5663	0.4381	0.4947	0.4858	0.4692
σ_2		τ=0.375		τ=0.5		τ=0.625	
κ_1	κ_2	score$_{PFP}$	score$_{FLR}$	score$_{PFP}$	score$_{FLR}$	score$_{PFP}$	score$_{FLR}$
1.8	0.0	0.7878	1.0331	0.7348	1.0428	0.8324	0.9867
0.8	0.0	0.8122	0.8565	0.7519	0.8338	0.8218	0.8461
0.2	0.0	0.7890	0.7795	0.7477	0.7346	0.7866	0.7812
0.4	0.4	0.8006	0.7913	0.7566	0.7558	0.8005	0.8037
0.0	0.8	0.8389	0.8349	0.7752	0.7974	0.8606	0.8340

Table 5: Mean width of the bootstrap prediction interval for different choices of κ_1 and κ_2 from 1000 bootstrap iterations

5 Real Data Analysis

5.1 PM10 concentration

One goal of this paper is to analyze the concentration of PM10 which broadly refers to particulate matter with an aerodynamic diameter of less than $10\mu m$ in ambient air, measured every 30 minutes in Graz, Austria. Before applying the proposed prediction method, we segmented the data vector according to the day of the week, then the 48 observations for each day were combined into a vector. Visual inspection of the data revealed several extreme outliers around New Year's Eve known to be caused by firework activities. The corresponding week is removed from the sample. Then we transformed the discrete vectors into functional objects with a 10-element cubic B-spline basis. Note that, in principle, the prediction results are anticipated to be robust to the choice of basis functions. Here, we have 175 daily functional observations. We also removed the daily mean for each day of the week to centralize the curves. To stabilize the variance, we performed the square root transformation. Figure 2 shows the trajectories after the preprocessing steps.

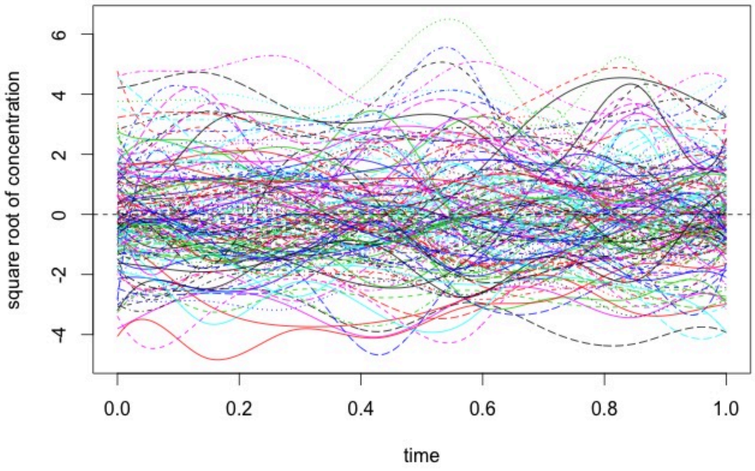

Figure 2: Centered square root transformed PM10 concentration curves

5.1.1 Prediction of smoothed PM10 concentration

We assumed the current time in a day is τ, where $\tau \in [0,1]$. Thus we have partial observation of the curve (i.e., we observe only on $[0,\tau)$ and thus needed to predict the curve over $[\tau,1]$. We used 87 curves to obtain the one-step ahead time series prediction in a sliding window way. Thus, there were 88 residual functions, among which the first 79 residual functions were used for estimating a fully functional regression model to update the prediction of the $[\tau,1]$ part of the rest curves. The one-step ahead prediction was conducted and the corresponding fFPE was computed. The averaged fFPE values are shown in Table 6 according to different values of p and d. Figure 3 shows the updated prediction of two randomly selected curve for various values of $\tau = 1/3, 1/2, 2/3$ respectively. In contrast to time series prediction methods and intraday regression method, PFP is superior with respect to the ℓ^2 prediction error of the unobserved part. Note that the prediction residual functions are not necessarily centered at zero, and thus the mean has to be adjusted when computing the intraday prediction. The final prediction is

$$\hat{Y}_{n+1}|_{(\tau,1]} \approx \hat{\mu}|_{(\tau,1]} + \hat{\mu}_e|_{(\tau,1]} + \sum_h (\hat{\Phi}_h(Y_{n+1-h} - \hat{\mu}))|_{(\tau,1]} + \hat{\beta}(Y_{n+1}|_{[0,\tau]} - \hat{Y}_{n+1}|_{[0,\tau]} - \hat{\mu}_e|_{[0,\tau]}),$$

where $\hat{\mu}_e$ is the estimated mean function of the prediction residual functions.

$\tau = 0.375$	fFPE							
	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$	$d = 8$
$p = 0$	**0.5993**	**0.5993**	**0.5993**	**0.5993**	**0.5993**	**0.5993**	**0.5993**	**0.5993**
$p = 1$	0.6278	0.6380	0.6330	0.6494	0.6459	0.6452	0.6462	0.6635
$p = 2$	0.6349	0.6591	0.6568	0.6695	0.6742	0.6965	0.7659	0.7933
$p = 3$	0.6357	0.6739	0.6412	0.6520	0.6542	0.7130	0.7966	0.8346
$\tau = 0.500$	fFPE							
	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$	$d = 8$
$p = 0$	**0.4184**	**0.4184**	**0.4184**	**0.4184**	**0.4184**	**0.4184**	**0.4184**	**0.4184**
$p = 1$	0.4274	0.4344	0.4260	0.4498	0.4655	0.4605	0.4485	0.4417
$p = 2$	0.4292	0.4460	0.4515	0.4691	0.4933	0.4962	0.5263	0.5441
$p = 3$	0.4268	0.4610	0.4385	0.4636	0.4830	0.5045	0.5657	0.5774
$\tau = 0.625$	fFPE							
	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$	$d = 8$
$p = 0$	0.1446	0.1446	0.1446	0.1446	0.1446	0.1446	0.1446	0.1446
$p = 1$	0.1517	0.1494	**0.1431**	0.1472	0.1444	0.1447	0.1525	0.1514
$p = 2$	0.1519	0.1490	0.1436	0.1494	0.1453	0.1474	0.1717	0.1744
$p = 3$	0.1514	0.1510	0.1535	0.1625	0.1580	0.1675	0.2004	0.1925

Table 6: The average fFPE value for different values of the order and dimension, when $\tau = 0.375$, $d_x = 6$, $d_y = 9$, $p = 0$; when $\tau = 0.5$, $d_x = 7$, $d_y = 8$, $p = 0$; when $\tau = 0.625$, $d_x = 8$, $d_y = 8$, $p = 1$, $d = 3$.

5.1.2 Comparison with moving block method

Shang (2017) proposed a functional time series prediction method, called the moving block method, to update the prediction with switching τ. Let τ to be the current time up to which the curve to be predicted is observed, then the time support are moved forward by τ. In other words, the $[\tau, 1]$ block of the m-th curve is combined with the $[0, \tau]$ block of the $(m + 1)$th curve to form a new function. The new functions are a recombination of the original functional time series with the loss of the $[0, \tau]$ part of the first curve, which has trivial effect on the prediction. The time series method is then

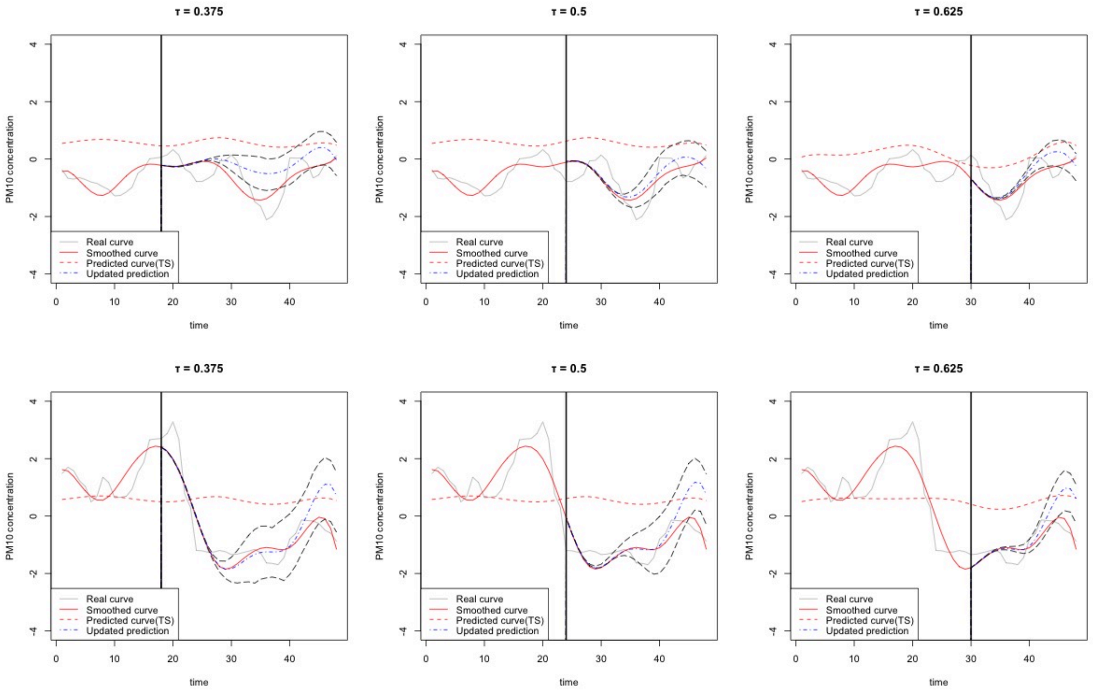

Figure 3: Updated predicted curve for different endpoints of the partially observed curve τ. The fitted curves (solid red line), the predicted curve by time series method (dotted red curve) and the predicted curve by PFP (dotted blue line after τ) with 95% bootstrap prediction intervals (upper and lower bound are shown by dotted black line) for a partially observed curve available up to $\tau = 1/3,\ 1/2,\ 2/3$, superimposed on the complete trajectory (gray line).

applied to the new functional time series, and the $[0, \tau]$ block of the predicted function is the update.

Table 7 gives the average of prediction MSE of the last 10 curves by PFP and the moving block method. It is noted that PFP robustly outperforms the moving block method over a broad range of values of τ. The result is not unexpected since the moving block method actually belongs to "next-interval" prediction method, which provides complete curve prediction, while PFP aims to produce prediction for the unobserved block, so the prediction error of the unobserved block of the new method should be smaller than that of the moving block method. Indeed one of the advantage of PFP over the moving block is that it directly uses the intraday variation, that is, the partially observed trajectory is directly treated as a part of the trajectory of interest rather than treating it artificially as part of the "previous" curve. A severe limitation of the moving block is that the partially observed curve for the current trajectory is artificially forced to be a part of the previous curve. This principle might work for some data settings but will not be reasonable for other many biological settings where the start of the curve has a well defined meaning (such as the onset of a stimulus presentation or a shock in biological experiments).

	τ=0.375	τ=0.5	τ=0.625
moving block	0.56194	0.34591	0.20138
new method	**0.34789**	**0.26852**	**0.10722**

Table 7: Prediction MSE of the two methods.

5.1.3 Prediction of the original curves

Since the PM10 curves are not smooth and present seasonal dynamics, it is natural to implement our method for the noisy case. The prediction result of PFP in the noisy case is compared with ARIMA model prediction, and PFP for smooth case is also implemented for comparison. We also applied linear interpolation when smoothing the original trajectories to incorporate the random error, and then used PFP in smooth case to finalize the prediction.

The current time τ was assumed to be 0.5, say the first 24 values were observed. The prediction methods are applied to predict the h-step ahead point values for the last 25 curves, where $1 \leq h \leq 10$. Table 8 shows the prediction error of the three methods. Figure 4 shows part of the original time series and the corresponding pre-smoothing residuals, and we note that after removing the smoothed functions, the residuals have no obvious seasonal behavior compared with the original one.

From Table 8, it is clear that there is dependence across the pre-smoothing residuals. Thus, by predicting the residuals, we expect significant improvement in the prediction.

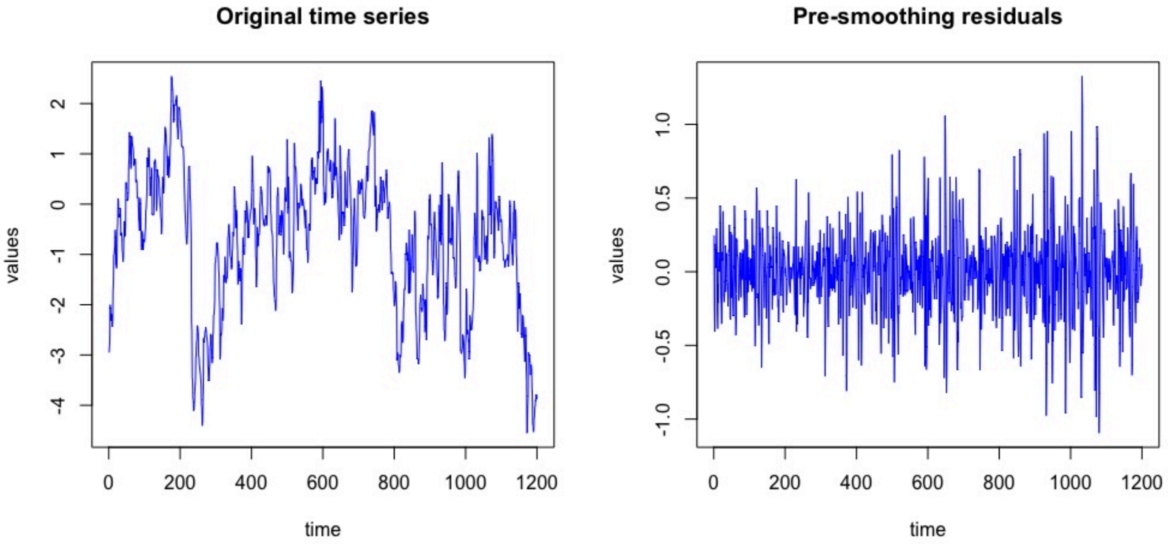

Figure 4: Part of the original time series and error time series

h	MSEc	MSEn	MSEa	MSEi
1	0.1508980	0.3540901	0.2307175	0.4642256
2	0.2680703	0.4757538	0.6123183	0.6500848
3	0.2309391	0.2576980	0.5273938	0.6132668
4	0.4849306	0.4972889	0.8758383	0.9419479
5	0.3512944	0.4108830	0.9005394	0.9464275
6	0.2363455	0.3411949	0.9953703	0.9732108
7	0.2317724	0.2619626	0.9681887	0.9455279
8	0.2184283	0.2406333	0.9389497	0.8657568
9	0.2376993	0.2154614	0.9931003	0.8339264
10	0.1853883	0.2210090	1.1429001	0.9116115

Table 8: Prediction MSE of the three methods. MSEc is the mean prediction MSE of PFP in the noisy case, MSEn is the mean prediction MSE of PFP in the smooth case, MSEa is the mean prediction MSE of ARIMA model, MSEi is the mean prediction error of PFP in the smooth case after linear interpolation.

These results demonstrate that PFP captures both the short-term dynamics (across pre-smoothing residuals) and long-term dynamics (across and within functions). The ARIMA model can only give good predictions for the short-term predicted values but will not give accurate predictions if we are interested in the long-term future. Linear interpolation does not perform well since the random error contaminates the smooth part and, thus, there could be bias in the principal components.

5.2 Traffic flow trajectories

We now analyze the traffic flow data that were collected by a dual loop vehicle detector near the Shea-San Tunnel on National Highway 5 in Taiwan in 2009 (shared by Chiou (2012)). It refers to the vehicle count per minute over 15-min time intervals (96 observations for each day). There are 92 days of observed trajectories in total, and the goal is to predict the unobserved block of the last 12 curves. In Figure 5, we show the raw daily trajectories and smoothed daily trajectories. Chiou (2012) proposed a functional mixture prediction method for independent trajectories. He first classified the trajectories into several clusters, and then used fully functional regression for intraday prediction of the unknown block in each potential cluster. The predictions in each cluster were combined to form the final prediction. It is obvious that the traffic flow trajectories had some specific patterns. Here, we used the first 80 curves as the training set to determine the cluster membership by subspace projection cluster algorithm (see Chiou and Li (2007)), and the last 12 curves are re-classified only based on the $[0, \tau]$ block. Intraday prediction is also conducted for comparison.

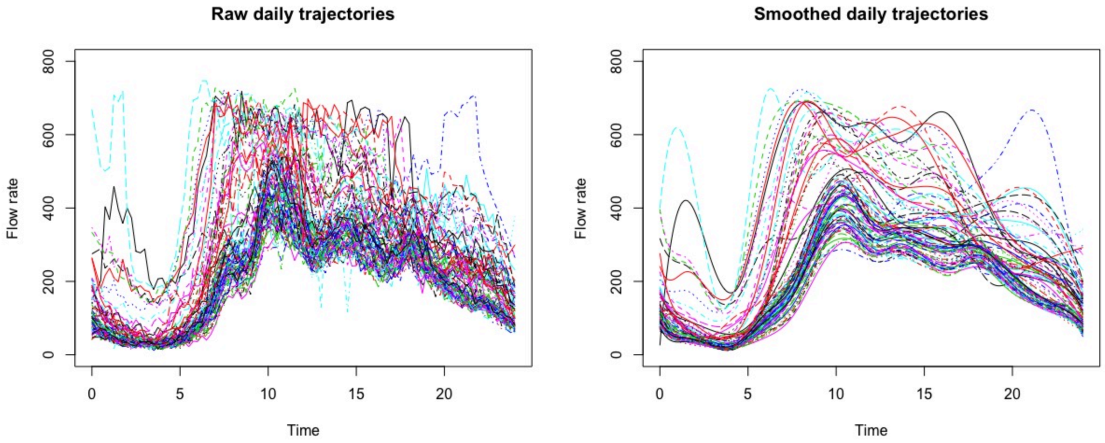

Figure 5: Daily traffic flow trajectories near theS hea-San Tunnel on National Highway 5 in Taiwan.

To demonstrate the necessity of time series structure, a comparative prediction was conducted. First we removed the daily mean for each day of the week to remove seasonal behavior. The window size for time series prediction is 40 curves, and the first 40 estimated innovation functions were used to predict the $[\tau, 1]$ block of the last 12 innovations.

In the test data, for a sample Y_i observed up to τ, we used the mean integrated prediction error (abbreviated as MIPE, see Chiou(2012)) to measure the performance of different methods. The MIPE can be expressed as

$$\text{MIPE}(\tau) = \frac{1}{12}\sum_{i=1}^{12}\frac{1}{1-\tau}\int_{\tau}^{1}\{Y_{i+80}(t)|_{(\tau,1]} - \hat{Y}_{i+80}(t)|_{(\tau,1]}\}^2 \mathrm{d}t.$$

Figure 6 shows the MIPE of the three methods.

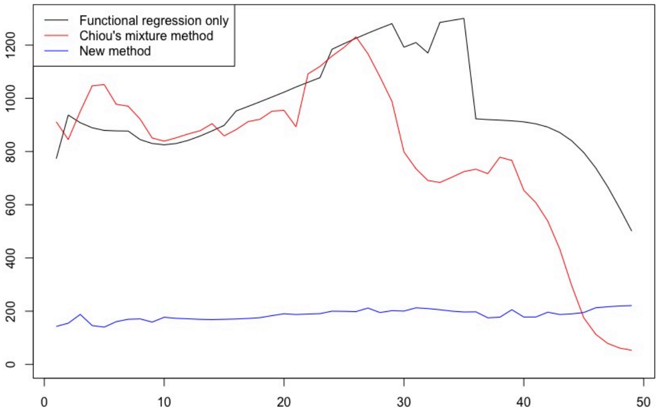

Figure 6: Integrated prediction error of the three methods corresponding to different τ ranging from 32 to 80, the index of the x-axis is $\tau_x - 31$, where τ_x is the index of the time grid up to which the curve is observed.

The result shows that PFP is superior compared with intraday prediction and functional mixture prediction method. In fact, functional mixture prediction method has some limitations. First, it requires that the curves has to be correctly classified and an incorrect membership could lead to poor results. Furthermore, applying FLR in each cluster actually reduces the sample size, and this will result in a larger estimation error. Another limitation is that the method classifies the future curve only based on the observed part, however, when the observed part is not very representative of the whole curve, the future curve to be predicted may be classified into a wrong cluster, which will potentially increase the prediction error.

6 Conclusion

This article proposes a new functional prediction methodology that provides an update on the prediction given that the curve to be predicted is partially observed. It is based on the idea that the updated prediction should be a projection onto the σ-algebra expanded by the past observed curves and the partial observation. The prediction algorithm is a stage-wise procedure, and can be applied to smooth and non-smooth functions. In non-smooth case, the functional techniques can be applied for removing the seasonal trend, and then ARMA model can be applied to predict the pre-smoothing residuals more effectively.

There are already a number of prediction methods for functions which we summarize here. In the functional time series prediction method (e.g. Aue et al. (2015)), the "next-interval" prediction only considers the big picture of the next function. In the setting where there is available partial observation, it is most natural to use the available data (in particular, intracurve information) to predict the unobserved part in order to improve prediction. The primary limitation of the existing functional time series methods is that they do not incorporate this available information. Another class of methods, the moving block method (Shang, 2017) is essentially it is "next-interval" prediction method, so it has the same limitations discussed above. Another limitation of this method is that it is unnatural to arbitrarily assign starting and ending points of a curve especially in studies where such are explicitly determined (e.g., start of the day; start of a trial in an experiment). For the fully functional regression method (see e.g. Ramsay and Silverman (2005)), while it is commended for incorporating intracurve information, its limitation is that it does not take into account the correlation across curves. This is a serious issue when the time series correlation is strong and that past curves are highly informative for predicting future curves. The method of Chiou et al. (2012) is an expansion of functional regression which very smartly and intuitively combines functional regression and clustering. The limitation of this approach is that when the partial observation does not give a strong indication of cluster membership, then the classification will not be reliable and this could lead to serious prediction errors. Indeed it can be challenging to classify time series with low signal to noise ratio and with short time series length, the classification may not be reliable. Moreover, the sample size is potentially reduced (per cluster) as we need to do estimation for each cluster separately. On the general approach of functional time series prediction after smoothing the curve by linear interpolation, the main difference between the general functional time series method is that this method smooths the curve by linear interpolation. This is a way to jointly incorporate the information of both long-term and short-term dynamics. But random errors will be included in the obtained curve and that could result in bias. Besides it is still “next-interval” prediction and thus also inherit all the limitations of the existing functional time series prediction methods.

PFP has several advantages. Since functional data are usually obtained in consecutive time intervals, the time series structure (e.g., autocorrelation) ubiquitously exists in functional data. PFP is the first one that takes time series into account for dynamic prediction update of functional data. The proposed fFPE criterion provides guidance to the user regarding whether or not the time series structure should be take into account when analyzing data. Thus, PFP is entirely data-driven. The simulation study and real data analysis demonstrates PFP always gives competitive prediction result.

References

[1] Antoniadis, A., Paparoditis, E, and Sapatinas, T. (2006). A Functional Wavelet-kernel Approach for Time Series Prediction. *Journal of Royal Statistical Society, Series B*, **68**, 837–857.

[2] Aue, A., Dubart Norinho, D., and Hörmann, S. (2015). On the Prediction of Stationary Functional Time series. *Journal of the American Statistical Association*, **110**, 378–392.

[3] Aue, A., Hörmann, S., Horváth, L., and Hušková, M. (2014). Dependent Functional Linear Models with Applications to Monitoring Structural Change. *Institute of Statistical Science, Academia Sinica*, **24**, 1043–1073.

[4] Besse, P.C., Cardot, H., and Stephenson, D.B. (2000). Autoregressive Forecasting of Some Functional Climate Variations. *The Scandinavian Journal of Statistics*, **27**, 673–687.

[5] Bosq, D. (2000). Linear Processes in Function Spaces, New York: Springer-Verlag.

[6] Chiou, J.-M. (2012). Dynamical Functional Prediction and Classification, with Application to Traffic Flow Prediction. *The Annals of Applied Statistics*, **6**, 1588–1614.

[7] Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of Royal Statistical Society, Series B*, **69**, 679–699.

[8] Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**:477, 359–378.

[9] Hörmann, S., Kokoszka, P. (2010). Weakly Dependent Functional Data. *The Annals of Statistics*, **38**, No. 3, 1845-1884

[10] Horváth, L., and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer, New York, Heidelberg, Dordrecht, London.

[11] Johnson, R.A., Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis, 6th Edition.* Pearson Education, Inc.

[12] Kargin, V., and Onatski, A. (2008). Curve Forecasting by Functional Autoregression. *Journal of Multivariate Analysis*, **99**, 2508–2526.

[13] Kokoszka. P., and Reimherr. M. (2013). Asymptotic normality of the principal componentes of functional time series. *Stochastic Process and their Application*, **123**, 1546–1562.

[14] Kokoszka, P., Reimherr, M. (2013). Determining the Order of the Functional Autoregressive Model. *Journal of Time series Analysis*, **34**, 116–129.

[15] Liu, X., Xiao, H., and Chen, R. (2016). Convolutional Autoregressive Models for Functional Time Series. *Journal of Econometrics*, **194**, 263–282.

[16] Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis.* Springer, Berlin Heidelberg, New York.

[17] Müller, H.G., Chiou, J.M., and Leng, X. (2008). Inferring Gene Expression Dynamics via Functional Regression Analysis. *BMC Bioinformatics*, **9**:60.

[18] Ramsay, J.O., and Silverman, B.W. (2005). *Functional Data Analysis, 2nd Edition.* Springer, New York.

[19] Shang, H.L. (2017), Functional Time Series Forecasting with Dynamic Updating: An Application to Intraday Particulate Matter Concentration. *Econometrics and Statistics*, **1**, 184–200.

[20] Shumway, R.H., and Stoffer, D.S. (2011). *Time Series Analysis and Its Application, 3rd Edition.* Springer, New York.

[21] Wang, J.L., Chiou, J.M., and Müller, H. G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Applications*, **3**, 257–295.

[22] Yao, F., Müller, H. G., and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Jornal of American Statistical Association*, **100**, 577–590.

Appendix

Theoretical proof

Lemma 1. Suppose $\{Y_k\} \in L^2[a,b]$ is an L^4-m approximable sequence, let $\{\phi_j(t),\ j \in \mathbb{Z}\}$ and $\{\hat{\phi}_j(t),\ j \in \mathbb{Z}\}$ be the eigenfunction and estimated eigenfunction respectively, and $\hat{c}_j = \text{sign}\langle \hat{\phi}_j, \phi_j\rangle$. Then we have $E[\langle \hat{c}_j\hat{\phi}_j, \phi_i\rangle] \to 0,$ for any pair of $i \neq j,$ and $E[\langle \hat{c}_j\hat{\phi}_j, \phi_j\rangle] \to 1,$ for any j.

Proof. When $i \neq j$, we have

$$E[|\langle \hat{c}_j\hat{\phi}_j, \phi_i\rangle|] = E[|\langle \hat{c}_j\hat{\phi}_j - \phi_j, \phi_i\rangle|] \leq \sqrt{E[\|\hat{c}_j\hat{\phi}_j - \phi_j\|^2]\|\phi_i\|^2}. \tag{A1}$$

Note that $\|\phi_i\| = 1$, and $E[\|\langle \hat{c}_j\hat{\phi}_j - \phi_j\rangle\|^2] \to 0$, then by Hörmann and Kokoszka (2010), then the right hand term in (A1) should converge to 0.

When $i = j$, we have

$$E[\langle \hat{c}_j\hat{\phi}_j, \phi_j\rangle] = E[\langle \hat{c}_j\hat{\phi}_j - \phi_j, \phi_j\rangle] + 1.$$

Similarly by Hörmann and Kokoszka (2010),

$$E[|\langle \hat{c}_j\hat{\phi}_j - \phi_j, \phi_j\rangle|] \leq \sqrt{E[|\hat{c}_j\hat{\phi}_j - \phi_j|^2]\|\phi_j\|^2} \to 0.$$

Thus $E[\langle \hat{c}_j\hat{\phi}_j, \phi_j\rangle] \to 1$.

Lemma 2. Suppose $\{Y_k\} \in L^2[a,b]$ is a functional sequence satisfying the condition in Lemma 1, with continuous covariance function, and $\{\xi_j,\ j \in \mathbb{Z}\}$ and $\{\hat{\xi}_j,\ j \in \mathbb{Z}\}$ be the eigenscore and estimated eigenscore respectively, and $\hat{c}_j = \text{sign}\langle\hat{\phi}_j, \phi_j\rangle$. Then we have

$$E[|\hat{c}_j\hat{\xi}_j - \xi_j|^2] \to 0, \text{for any } j \in \mathbb{Z}.$$

Furthermore, we have

$$E[\hat{c}_j\hat{\xi}_j\xi_i] \to 0,\ E[\hat{c}_j\hat{\xi}_j\hat{c}_i\hat{\xi}_i] \to 0,\ E[\hat{c}_j\hat{\xi}_j\xi_j] \to \lambda_j,\ E[\hat{\xi}_j^2] \to \lambda_j.$$

Proof. By Hölder inequality,

$$|\hat{c}_j\hat{\xi}_j - \xi_j|^2 = \langle Y, \hat{c}_j\hat{\phi}_j - \phi_j\rangle^2 \leq \|Y\|^2\|\hat{c}_j\hat{\phi}_j - \phi_j\|^2.$$

We can assume that $\hat{\phi}_j$ is obtained from an independent copy of the original sample, since correlation between the estimated covariance operator and a single functional sample Y should be negligible. Then we have

$$E[|\hat{c}_j\hat{\xi}_j - \xi_j|^2] \leq E[\|Y\|^2]E[\|\hat{c}_j\hat{\phi}_j - \phi_j\|^2].$$

We know that $E[\|Y\|^2] = \int_a^b C(t,t)dt < \infty$, and by Hörmann and Kokoszka (2010), $E[\|\hat{c}_j\hat{\phi}_j - \phi_j\|^2] \to 0$, thus $E[|\hat{c}_j\hat{\xi}_j - \xi_j|^2] \to 0$.

Then by the Mercer's theorem we can get $E[\hat{c}_j\hat{\xi}_j\xi_i] = E[(\hat{c}_j\hat{\xi}_j - \xi_j)\xi_i]$, using the result we have just obtained, we can get

$$E[|(\hat{c}_j\hat{\xi}_j - \xi_j)\xi_i|] \leq \sqrt{E[(\hat{c}_j\hat{\xi}_j - \xi_j)^2]E[\xi_i^2]} = \sqrt{\lambda_i E[(\hat{c}_j\hat{\xi}_j - \xi_j)^2]} \to 0.$$

Similarly, we have $E[\hat{c}_j\hat{\xi}_j\xi_j] = E[(\hat{c}_j\hat{\xi}_j - \xi_j)\xi_j] + \lambda_j$, and $E[|(\hat{c}_j\hat{\xi}_j - \xi_j)\xi_j|] < \sqrt{E[(\hat{c}_j\hat{\xi}_j - \xi_j)^2]E[\xi_j^2]} \to 0$, thus $E[\hat{c}_j\hat{\xi}_j\xi_j] \to \lambda_j$.

We also have $E[\hat{c}_j\hat{\xi}_j\hat{c}_i\hat{\xi}_i] = E[\hat{c}_j\hat{\xi}_j\hat{c}_i\hat{\xi}_i - \hat{c}_j\hat{\xi}_j\xi_i + \hat{c}_j\hat{\xi}_j\xi_i - \xi_j\xi_i]$, and by Cauchy-Schwarz inequality,

$$\begin{aligned} E[|\hat{c}_j\hat{\xi}_j\hat{c}_i\hat{\xi}_i - \hat{c}_j\hat{\xi}_j\xi_i + \hat{c}_j\hat{\xi}_j\xi_i - \xi_j\xi_i|] &\leq E[|\hat{c}_j\hat{\xi}_j(\hat{c}_i\hat{\xi}_i - \xi_i)|] + E[|(\hat{c}_j\hat{\xi}_j - \xi_j)\xi_i|] \\ &\leq \sqrt{E[\hat{\xi}_j^2]E[(\hat{c}_i\hat{\xi}_i - \xi_i)^2]} + \sqrt{E[\xi_j^2]E[(\hat{c}_i\hat{\xi}_i - \xi_i)^2]}. \end{aligned} \tag{A2}$$

Since $\hat{c}_j\hat{\xi}_j \xrightarrow{m.s.} \xi_j$, so $E[\hat{\xi}_j^2]$ must be bounded. Then (A2) converge to zero. Since $E[\hat{\xi}_j] \xrightarrow{m.s.} E[\xi_j]$, so $E[\hat{c}_j\hat{\xi}_j^2] \to E[\xi_j^2] = \lambda_j$.

Proof of Theorem 1. It is obvious that

$$\begin{aligned} E[\|Y_{n+1} - \hat{Y}_{n+1}\|^2] &= E[\|Y_{n+1} - \tilde{Y}_{n+1} + \tilde{Y}_{n+1} - \hat{Y}_{n+1}\|^2] \\ &= E[\|Y_{n+1} - \tilde{Y}_{n+1}\|^2] + E[\|\tilde{Y}_{n+1} - \hat{Y}_{n+1}\|^2] + 2E[\langle Y_{n+1} - \tilde{Y}_{n+1}, \tilde{Y}_{n+1} - \hat{Y}_{n+1}\rangle]. \end{aligned}$$

Thus it is suffice to show $E[\|\tilde{Y}_{n+1} - \hat{Y}_{n+1}\|^2] \to 0$ and $E[\langle Y_{n+1} - \tilde{Y}_{n+1}, \tilde{Y}_{n+1} - \hat{Y}_{n+1}\rangle] \to 0$.

First we need to show $E[\|\tilde{Y}_{n+1} - \hat{Y}_{n+1}\|^2] \to 0$. Assume $\hat{\beta}_{ij}$ is the estimation of β_{ij} based on real eigenscores, and $\tilde{\beta}_{ij}$ is the estimation of β_{ij} based on real eigenscores. Then we have

$$\begin{aligned}
E[\|\tilde{Y}_{n+1} - \hat{Y}_{n+1}\|^2] &= E[\| \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right) \hat{d}_j \hat{\psi}_j(t) - \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \hat{\beta}_{ij} \xi_i \right) \psi_j(t) \|^2] \\
&= E[\| \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right) \hat{d}_j \hat{\psi}_j(t) - \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right) \psi_j(t) \\
&\qquad + \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right) \psi_j(t) - \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \hat{\beta}_{ij} \xi_i \right) \psi_j(t) \|^2] \\
&\leq 2E[\| \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right) \hat{d}_j \hat{\psi}_j(t) - \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right) \psi_j(t) \|^2] \quad \text{(A3)} \\
&\qquad + 2E[\| \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right) \psi_j(t) - \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \hat{\beta}_{ij} \xi_i \right) \psi_j(t) \|^2]. \quad \text{(A4)}
\end{aligned}$$

First we need to (A3) converge to zero. By the inequality $\| \sum_{k=1}^{m} a_k \|^2 \leq m \sum_{k=1}^{m} \|a_k\|^2$, we have

$$E\left[\| \sum_{j=1}^{d_y} \left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right) \left(\hat{d}_j \hat{\psi}_j(t) - \psi_j(t) \right) \|^2 \right] \leq d_y \sum_{j=1}^{d_y} E\left[\left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right)^2 \|\hat{d}_j \hat{\psi}_j(t) - \psi_j(t)\|^2 \right].$$

By Theorem 1 in Aue et al. (2014), we have $\tilde{\beta} = \beta + O_p(n^{-1/2})$, so $\tilde{\beta} \xrightarrow{p} \beta$, by Lemma 1 $\hat{c}_i \hat{\xi}_i \xrightarrow{p} \xi_i$ and by Hörmann and Kokoszka (2010), $\|\hat{d}_j \hat{\psi}_j(t) - \psi_j(t)\|^2 \xrightarrow{p} 0$. So by continuous mapping theorem,

$$\left(\sum_{i=1}^{d_x} \tilde{\beta}_{ij} \hat{c}_i \hat{\xi}_i \right)^2 \|\hat{d}_j \hat{\psi}_j(t) - \psi_j(t)\|^2 \xrightarrow{p} 0. \quad \text{(A5)}$$

We have $E\left[\|\hat{d}_j \hat{\psi}_j(t) - \psi_j(t)\|^{4+\epsilon} \right] < \infty$ if $E[Y^4(t) \otimes Y^4(s)] < \infty$. By Hormann and Kokoszka (2010),

$$\|\hat{d}_j \hat{\psi}_j - \psi_j\| \leq \frac{2\sqrt{2}}{\alpha_j} \|\hat{C} - C\|_{\mathcal{L}} \leq \frac{2\sqrt{2}}{\alpha_j} \|\hat{C} - C\|_{\mathcal{S}},$$

where $\alpha_1 = \lambda_1 - \lambda_2$ and $\alpha_j = \min\{\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}\}$, $j \geq 2$. In order to make $E\left[\|\hat{d}_j\hat{\psi}_j(t) - \psi_j(t)\|^{4+\epsilon}\right] < \infty$ holds, we only need $E\|\hat{C} - C\|_{\mathcal{S}}^{4+\epsilon} < \infty$. And we have

$$\begin{aligned} E\|\hat{C} - C\|_{\mathcal{S}}^{4+\epsilon} &= E\left(\int\int\left[\frac{1}{n}\sum_{k=1}^{n}(Y_k(t)Y_k(s) - E[Y_k(t)Y_k(s)])\right]^2 dtds\right)^{2+\frac{\epsilon}{2}} \\ &\leq c_\tau\left(\int\int E\left[\frac{1}{n}\sum_{k=1}^{n}(Y_k(t)Y_k(s) - E[Y_k(t)Y_k(s)])\right]^4 dtds\right)^{1+\frac{\epsilon}{4}} < \infty, \end{aligned}$$

where c_τ is a constant that is related to τ since the integration is taken on a closed interval.

And we can assume that $\tilde{\beta}_{ij}$ is obtained from an independent copy of X, Y's, then $\tilde{\beta}_{ij}$ should be independent with $\hat{\xi}_i$, and $E[\hat{\xi}_i^{4+\epsilon}] = E[\int X\phi_i dt]^{4+\epsilon} \leq E[\|X\|^{4+\epsilon}]\|\phi_i\|^{4+\epsilon} = E[\|X\|^{4+\epsilon}\|] < \infty$, and $\tilde{\beta}_{ij}$ is asymptoticly normal, so $E[\tilde{\beta}_{ij}^{4+\epsilon}] < \infty$. Then we have $E\left[\left(\sum_{i=1}^{d_x}\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i\right)^{4+\epsilon}\right] < \infty$.

By Cauchy-Schwarz inequality and the assumption that $E[\|X_k\|^{4+\epsilon}] < \infty$ and $E[\|Y_k\|^{4+\epsilon}] < \infty$, we have

$$E\left[\left(\sum_{i=1}^{d_x}\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i\right)^2\|\hat{d}_j\hat{\psi}_j(t) - \psi_j(t)\|^2\right]^{1+\epsilon'} \leq \sqrt{E\left[\left(\sum_{i=1}^{d_x}\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i\right)^{4+\epsilon}\right]E\left[\|\hat{d}_j\hat{\psi}_j(t) - \psi_j(t)\|^{4+\epsilon}\right]} < \infty,$$

where $\epsilon' = \epsilon/4$. Thus the term in (A5) is uniformly integrable, then we have

$$E\left[\left(\sum_{i=1}^{d_x}\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i\right)^2\|\hat{d}_j\hat{\psi}_j(t) - \psi_j(t)\|^2\right] \to 0.$$

As for the second term (A4), we have

$$E\left[\|\sum_{j=1}^{d_y}\left(\sum_{i=1}^{d_x}\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i-\sum_{i=1}^{d_x}\hat{\beta}_{ij}\xi_i\right)\psi_j(t)\|^2\right]<d_xd_y\sum_{j=1}^{d_y}\sum_{i=1}^{d_x}E\left[\left(\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i-\hat{\beta}_{ij}\xi_i\right)^2\right].$$

We need to show $E\left[\left(\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i-\hat{\beta}_{ij}\xi_i\right)^2\right]$ converge to zero. Under the assumption that $\tilde{\beta}$ and $\hat{\beta}$ are obtained from an independent copy of the sample $(Y_k\colon k\in\mathbb{N})$, it is clear that

$$\begin{aligned}
E\left[\left(\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i-\hat{\beta}_{ij}\xi_i\right)^2\right] &= E\left[\left(\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i-\tilde{\beta}_{ij}\xi_i+\tilde{\beta}_{ij}\xi_i-\hat{\beta}_{ij}\xi_i\right)^2\right]\\
&< 2\left\{E\left[\left(\tilde{\beta}_{ij}\hat{c}_i\hat{\xi}_i-\tilde{\beta}_{ij}\xi_i\right)^2\right]+E\left[\left(\tilde{\beta}_{ij}\xi_i-\hat{\beta}_{ij}\xi_i\right)^2\right]\right\}\\
&= 2\left\{E\left[\tilde{\beta}_{ij}^2\right]E\left[\left(\hat{c}_i\hat{\xi}_i-\xi_i\right)^2\right]+E\left[\left(\tilde{\beta}_{ij}-\hat{\beta}_{ij}\right)^2\right]E\left[\xi_i^2\right]\right\},
\end{aligned}$$

which should converge to zero by Lemma 2 and Theorem 1 in Aue et al. (2014). In fact,

$$E\left[\left(\tilde{\beta}_{ij}-\hat{\beta}_{ij}\right)^2\right]\leq 2\left\{E\left[\left(\tilde{\beta}_{ij}-\beta_{ij}\right)^2\right]+E\left[\left(\beta_{ij}-\hat{\beta}_{ij}\right)^2\right]\right\},$$

and we have already shown the second term converge to zero. As for the first term, by Theorem 1 in Aue et al. (2014) and Kokoszka et al. (2013) we can prove it.

To finish the proof, we only need to show that $E[\langle Y_{n+1}-\tilde{Y}_{n+1},\tilde{Y}_{n+1}-\hat{Y}_{n+1}\rangle]$ converges to zero. It is clear that

$$E[\langle Y_{n+1}-\tilde{Y}_{n+1},\tilde{Y}_{n+1}-\hat{Y}_{n+1}\rangle]=E[\langle Y_{n+1}-\hat{Y}_{n+1},\tilde{Y}_{n+1}-\hat{Y}_{n+1}\rangle]+E[\|\tilde{Y}_{n+1}-\hat{Y}_{n+1}\|^2],$$

then it suffice to show $E[\langle Y_{n+1}-\hat{Y}_{n+1},\tilde{Y}_{n+1}-\hat{Y}_{n+1}\rangle]\to 0$. By Cauchy-Schwarz inequality,

$$E[|\langle Y_{n+1}-\hat{Y}_{n+1},\tilde{Y}_{n+1}-\hat{Y}_{n+1}\rangle|]\leq\sqrt{E[\|Y_{n+1}-\hat{Y}_{n+1}\|]^2\times E[\|\tilde{Y}_{n+1}-\hat{Y}_{n+1}\|^2]}.$$

For a well defined regression model, expected prediction mean square error should be finite, and by the previous result, we have $E[\langle Y_{n+1} - \hat{Y}_{n+1}, \tilde{Y}_{n+1} - \hat{Y}_{n+1} \rangle] \to 0$.

Additional figures

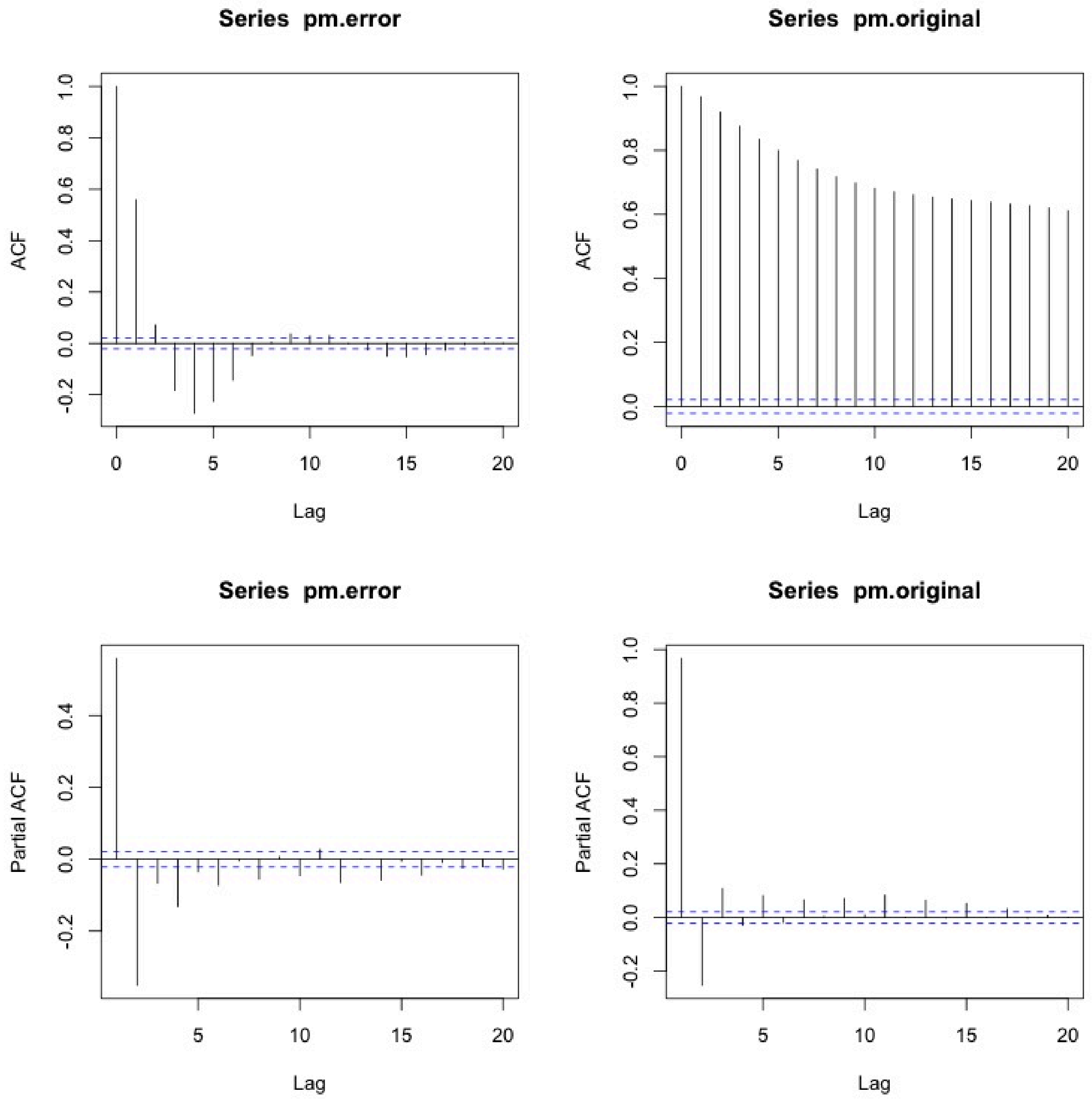

Figure 7: The autocorrelation function and partial autocorrelation function of the pre-smoothing residuals and the original PM10 concentration process.