# Functional Autoregressive Processes on Multidimensional Data

Qin Wen

Advisors: Rebecca Willett, Daren Wang

## Abstract

We study the inference (estimation and prediction) of a functional autoregressive (FAR) process, a statistical tool for modeling functional time series data. We proposed a Frobenius norm regularization method to estimate the transition operator of the FAR process directly from discrete measurements of the functional time series. We extend the one-dimensional framework via the tools of Reproducing Kernel Hilbert Spaces (RKHS method) into two- dimensional scenario. We also developed the two-dimensional prediction method based on functional principal component analysis (FPCA) which serves as a baseline method in this paper. Extensive numerical results demonstrated the RKHS method is significantly better when recovering the operator.

# Contents

# 1 Introduction

Functional data analysis (FDA) deal with data in the form of curves, surfaces or anything else varying over a continuum. As described in Ramsay and Silverman (2005), Ferraty and Vieu (2006) and Horváth and Kokoszka (2012), FDA framework considers each sample element to be a function, offering a natural solution to a variety of problems which are not suitable for the classical statistical frameworks.

When the functional data is observed in a sequential manner, it is called functional time series (Hörmann and Kokoszka, 2010; Aue et al., 2015). Formally speaking, a functional time series takes the form $\{X_t(s), s \in [a, b]\}_{t\in\mathbb{Z}}$, where each observation $X_t(s)$ is a (random) function defined for $s$ taking values in some compact interval $[a, b]$. Throughout the paper, we assume that $[a, b] = [0, 1]$. It may well happen that the classical vector autoregressive models would fail to track the dynamics of functional time series in high-dimensional spaces. (Bosq (2000), Hyndman and Ullah (2007) and Shang (2013)). Therefore, an important goal for functional time series analysis is the estimation and prediction of a reliable statistical model, which will offer a better understanding of the behavior of the data and providing accurate prediction for future realization.

In this paper, we are going to talk about a widely-used functional time series model: functional autoregressive(FAR) process. Formally speaking, a functional time series $\{X_t\}$ follows an FAR process (in $\mathcal{L}^2$) of order $D$ if

$$X_t(\cdot) = \mu(\cdot) + \sum_{d=1}^{D} \Psi_d(X_{t-d})(\cdot) + \epsilon_t(\cdot), \tag{1}$$

where $\mu(\cdot) \in \mathcal{L}^2$ is a deterministic function, $\epsilon_t(\cdot) \in \mathcal{L}^2$ are *i.i.d.* zero-mean noise functions, and $\{\Psi_d\}_{d=1}^{D}$ are bounded linear operators mapping $\mathcal{L}^2 \to \mathcal{L}^2$. An important goal of this time series analysis is providing reliable predictions for future realizations.

Dimension reduction is mainly used to address the inference of FAR in the existing literature due to the infinite-dimensional nature of the functional space $\mathcal{L}^2$. Most literature reduces the dimension via functional principal component analysis (FPCA), where the finite basis is chosen as the leading $p$ functional principal components (FPC) of an estimated covariance operator of the underlying FAR process. More works on FPCA-based approaches can be found at Besse and Cardot (1996), Bosq (2000), Besse et al. (2000), Hyndman and Shang (2009), Didericksen et al. (2012) and Aue et al. (2015).

Existing estimation methods and theoretical results require fully observed functional time series $\{X_t(s), s \in [0,1]\}_{t=1}^T$. However, in reality, this is an unrealistic assumption as the FAR process is measured discretely and observations instead take the form $\{X_t(s_i), 1 \leq i \leq n\}_{t=1}^T$, where $\{s_i\}_{i=1}^n$ denotes $n$ discrete grid points in $[0,1]$. In practice, the FPCA-based methods typically rely on an extra smoothing step to convert discrete measurements $\{X_t(s_i), 1 \leq i \leq n\}_{t=1}^T$ into (estimated) fully functional data $\{\widetilde{X}_t(s), s \in [0,1]\}_{t=1}^T$, and the statistical analysis is performed on $\{\widetilde{X}_t(s), s \in [0,1]\}_{t=1}^T$. This will bring the smoothing error when establishing theoretical guarantees.

Wang et al. (2020a) proposed new estimation and prediction procedures for FAR without dimension reduction through the lens of Reproducing Kernel Hilbert Spaces (RKHS, Wahba (1990)). To be more specific, the paper considered a refined FAR process in RKHS,

$$X_t(\cdot) = \mu(\cdot) + \sum_{d=1}^{D} \int_0^1 A_d(\cdot, s) X_{t-d}(s) ds + \epsilon_t(\cdot), \tag{2}$$

where the bounded linear operators $\{\Psi_d\}_{d=1}^D$ take the explicit form of integral operators with bivariate kernels $\{A_d(r,s) : [0,1] \times [0,1] \to \mathbb{R}\}_{d=1}^D$. The paper construct theoretical guarantee for estimation and prediction of FAR processes based on discrete observations of functional time series.

A limitation of this paper is that they used the accelerated gradient method(AGM) to solve the optimization problem. Notice that this algorithm does singular value decomposition in every iteration, which will be time-consuming, especially when we would like to apply this framework into multi-dimension. Therefore, in our paper, we explored the complexity and regularization of the regression problem to develop a computational efficient algorithm, and extended the 1 dimensional framework into multi-dimension. We also extended the fPCA method into 2 dimensional scenario, which serves as a baseline method and applied them to our problem.

The rest of the paper is organized as follows. Section 2 proposes the penalized Frobenius norm estimator for FAR and studies its theoretical properties. The performance of the 1 dimensional proposed method over existing procedures is demonstrated via extensive numerical experiments in Section 3. Section 4 extend both methods based on regression and based on fPCA into 2-dimension. The performance of 2 dimensional methods can be found in section 5.

Some notations used throughout the paper are defined as follows. Denote $\|f\|_{\mathcal{L}^2}^2 = \langle f, f\rangle_{\mathcal{L}^2}$ and $\|f\|_\infty := \sup_{s\in[0,1]} |f(s)|$. For a matrix $W$, denote $\|W\|_F$ as its Frobenius norm and $\|W\|_*$ as its trace norm. We omit $[0,1]$ in the integral whenever the domain of functions is clear.

# 2 Estimation Methodology and Main Results

Section 2.1 proposes the RKHS-based penalized estimation procedure for the transition operators $\{A_d^*\}_{d=1}^D$ with discrete realizations of $\{X_t\}_{t=1}^T$. Section 2.2 formulates the penalized estimation as a Frobenius norm minimization problem rather than nuclear norm as in Wang et al. (2020a) and discusses its numerical implementation.

Let us define the FAR(D) process in RKHS.

**Definition 1.** *A functional time series $\{X_t\}_{t=1}^T \subset \mathcal{H}$ is said to follow a functional autoregressive process of order $D$ in $\mathcal{H}$(FAR(D)) for an RKHS $\mathcal{H}$, if*

$$X_t(r) = \sum_{d=1}^D \int A_d^*(r,s)X_{t-d}(s)ds + \epsilon_t(r), \text{ for } r \in [0,1], \tag{3}$$

*where $\{\epsilon_t\}_{t=1}^T \subset \mathcal{H}$ is a collection of i.i.d. functional noise and the transition operators $\{A_d^*\}_{d=1}^D \subset \mathcal{C}$ are compact linear operators on $\mathcal{H}$.*

## 2.1 Penalized estimation and Representer theorem

Existing literature requires fully observed functional time series, but this is impractical because the FAR process is measured discretely in reality. Following the standard RKHS literature, we assume $\{s_i\}_{i=1}^n$ to be a collection of random variables uniformly sampled from

$[0,1]$. Given the discrete observations $\{X_t(s_i)\}_{1\le t\le T, 1\le i\le n}$, our goal is to estimate the $D$ unknown transition operators $\{A_d^*\}_{d=1}^D$, which will facilitate the prediction task.

We construct the estimator $\{\widehat{A}_d\}_{d=1}^D$ through a penalized Frobenius norm optimization such that

$$\{\widehat{A}_d\}_{d=1}^D = \mathop{\arg\min}_{\{A_d\}_{d=1}^D \in \mathcal{C}_{\boldsymbol{\tau}}} \frac{1}{Tn} \sum_{t=D+1}^{T} \sum_{i=1}^{n} \left( X_t(s_i) - \sum_{d=1}^{D} \frac{1}{n} \sum_{j=1}^{n} A_d(s_i, s_j) X_{t-d}(s_j) \right)^2 \tag{4}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_D)$ is the tuning parameter and $\mathcal{C}_{\boldsymbol{\tau}} := \{(A_1, \cdots, A_D) : A_d \in \mathcal{C}$ and $\|A_d\|_{\mathcal{H},\mathcal{F}} \le \tau_d,\ d = 1, \cdots, D\}$ is the constraint space. We name $\{\widehat{A}_d\}_{d=1}^D$ in (4) the constrained Frobenius norm estimator for transition operators of FAR.

The motivation for the formulation of (4), is that consider the integral scenario (ideal yet infeasible) in which $\{X_t\}_{t=1}^T$ are fully observed in the entire domain $[0,1]$, thus we can solve

$$\{\widetilde{A}_d\}_{d=1}^D = \mathop{\arg\min}_{\{A_d\}_{d=1}^D \in \mathcal{C}_{\boldsymbol{\tau}}} \frac{1}{T} \sum_{t=D+1}^{T} \int \left( X_t(r) - \sum_{d=1}^{D} \int A_d(r,s) X_{t-d}(s) ds \right)^2 dr.$$

After using the integral approximation $\int A_d(s_i, r) X_{t-d}(r) dr \approx \frac{1}{n} \sum_{j=1}^n A_d(s_i, s_j) X_{t-d}(s_j)$, we instead solve (4) since only discrete data $\{X_t(s_i)\}_{1\le t\le T, 1\le i\le n}$ are observed.

**Proposition 1** (Representer theorem)**.** *There exists a minimizer $\{\widehat{A}_d\}_{d=1}^D$ of the constrained Frobenius/nuclear norm optimization (4) such that for any $(r,s) \in [0,1] \times [0,1]$,*

$$\widehat{A}_d(r,s) = \sum_{1\le i,j\le n} \widehat{a}_{d,ij} \mathbb{K}(r, s_i) \mathbb{K}(s, s_j), \ \textit{for } d = 1, 2, \cdots, D. \tag{5}$$

Wang et al. (2020a) propose a representer theorem of the nuclear norm optimization, whose proof is also suitable to establish the Representer theorem for Frobenius norm optimization.

After obtaining the estimated transition operators $\{\widehat{A}_d\}_{d=1}^D$, the one-step ahead prediction of $X_{T+1}$ can be computed as

$$\widehat{X}_{T+1}(r) = \sum_{d=1}^{D} \frac{1}{n} \sum_{j=1}^{n} \widehat{A}_d(r, s_j) X_{T+1-d}(s_j) \text{ for } r \in [0,1]. \tag{6}$$

## 2.2 Optimization for One-dimensional Scenario

In this section, we discuss the numerical implementation of the proposed RKHS-based penalized estimator by reformulating the constrained optimization in (4) into a Frobenius norm minimization problem.

We first introduce some notations. Denote the estimator $A_d(r,s) = \sum_{1\le i,j\le n} a_{d,ij}\mathbb{K}(r,s_i)\mathbb{K}(s,s_j)$, where $a_{d,ij}$s are the coefficients to be estimated. Define the coefficient matrix $R_d \in \mathbb{R}^{n\times n}$ with $R_{d,ij} = a_{d,ij}$. Define the kernel vector $k_i = (\mathbb{K}(s_1,s_i), \mathbb{K}(s_2,s_i), \cdots, \mathbb{K}(s_n,s_i))^\top$ and the kernel matrix $K = [k_1, k_2, \cdots, k_n]$. Note that the kernel matrix $K$ is symmetric such that $K = K^\top$. Denote the observation of the functional time series at time $t$ as $X_t = (X_t(s_1), X_t(s_2), \cdots, X_t(s_n))^\top$. Define the observation matrix $X = [X_T, X_{T-1}, \cdots, X_{D+1}]$ and the lagged observation matrix $X^{(d)} = [X_{T-d}, X_{T-d-1}, \cdots, X_{D+1-d}]$ for $d = 1, \cdots, D$.

Using the well-known equivalence between constrained and penalized optimization (see Hastie et al. (2009)), we can reformulate (4) into a penalized Frobenius norm optimization such that

$$\{\widehat{A}_d\}_{d=1}^D = \arg\min \sum_{t=D+1}^{T}\sum_{i=1}^{n}\left(X_t(s_i) - \sum_{d=1}^{D}\frac{1}{n}\sum_{j=1}^{n}A_d(s_i,s_j)X_{t-d}(s_j)\right)^2 + \sum_{d=1}^{D}\lambda_d\|A_d\|_{\mathcal{H},\mathcal{F}}^2,$$

where $(\lambda_1, \cdots, \lambda_D)$ is the tuning parameter. With simple linear algebra, we can rewrite the penalized optimization as

$$\begin{aligned}&\min_{R_1,\cdots,R_D}\sum_{t=D+1}^{T}\left(X_t - \frac{1}{n}\sum_{d=1}^{D}K^\top R_d K X_{t-d}\right)^\top\left(X_t - \frac{1}{n}\sum_{d=1}^{D}K^\top R_d K X_{t-d}\right) + \sum_{d=1}^{D}\lambda_d\|A_d\|_{\mathcal{H},\mathcal{F}}^2\\ =&\min_{R_1,\cdots,R_D}\left\|X - \frac{1}{n}\sum_{d=1}^{D}KR_dKX^{(d)}\right\|_F^2 + \sum_{d=1}^{D}\lambda_d\|A_d\|_{\mathcal{H},\mathcal{F}}^2. \qquad (7)\end{aligned}$$

We now write the Frobenius norm $\|A_d\|_{\mathcal{H},\mathcal{F}}$ as a function of $R_d$. By the Representer theorem, $A_d(r,s) = \sum_{i,j} a_{d,ij}\mathbb{K}(r,s_i)\mathbb{K}(s,s_j)$, thus the adjoint operator $A_d^\top(r,s) = A_d(s,r)$. Define $k(s) = (\mathbb{K}(s,s_1), \mathbb{K}(s,s_2), \cdots, \mathbb{K}(s,s_n))^\top$, we have $A_d(r,s) = k(r)^\top R_d k(s)$ and $\langle k(s), k(s)^\top\rangle_{\mathcal{H}} =$

$K$. Define $u(s) = k(s)^\top b$, where $b = (b_1, b_2, \cdots, b_n)^\top$. To calculate $\|A_d\|_{\mathcal{H},\mathcal{F}}$, note that

$$\begin{aligned} A_d^\top A_d[u](s) &= \langle A_d^\top(s,r), A_d[u](r)\rangle_{\mathcal{H}} = \langle A_d(r,s), \langle A_d(r,s), u(s)\rangle_{\mathcal{H}}\rangle_{\mathcal{H}} \\ &= \langle k(r)^\top R_d k(s), \langle k(r)^\top R_d k(s), k(s)^\top b\rangle_{\mathcal{H}}\rangle_{\mathcal{H}} = k(s)^\top R_d^\top K R_d K b. \end{aligned}$$

In other words, the eigenvalues of the operator $A_d^\top A$ correspond to the eigenvalues of the matrix $R_d^\top K R_d K$. Thus, (7) can be further written as

$$\begin{aligned} &\min_{R_1,\cdots,R_D} \left\| X - \frac{1}{n}\sum_{d=1}^{D} K R_d K X^{(d)} \right\|_F^2 + \sum_{d=1}^{D} \lambda_d \cdot \mathrm{trace}((R_d^\top K R_d K)) \\ =&\min_{R_1,\cdots,R_D} \left\| X - \frac{1}{n}\sum_{d=1}^{D} K R_d K X^{(d)} \right\|_F^2 + \sum_{d=1}^{D} \lambda_d \|K^{\frac{1}{2}} R_d K^{\frac{1}{2}}\|_{\mathcal{F}}^2 \\ =&\min_{W_1,\cdots,W_D} \left\| X - \sum_{d=1}^{D} \mathcal{K} W_d Z_d \right\|_F^2 + \sum_{d=1}^{D} \|W_d\|_{\mathcal{F}}^2 \end{aligned} \tag{8}$$

where $W_d = K^{\frac{1}{2}} R_d K^{\frac{1}{2}}$, $\mathcal{K} = K^{\frac{1}{2}}$, $Z_d = \frac{1}{n} K^{\frac{1}{2}} X^{(d)}$ and the first equality comes from the fact that $R_d^\top K R_d K$ and $K^{1/2} R_d^\top K R_d K^{1/2}$ share the same eigenvalues for $d = 1, \cdots, D$.

In the following experiment, we set d = 1. Hence, the objective function becomes

$$\begin{aligned} &\min_R \left\| X - \frac{1}{n} K R K X^{(1)} \right\|_F^2 + \lambda \cdot \|K^{\frac{1}{2}} R K^{\frac{1}{2}}\|_{\mathcal{F}}^2 \\ =&\min_W \left\| X - \frac{1}{n} K^{\frac{1}{2}} W K^{\frac{1}{2}} X^{(1)} \right\|_F^2 + \lambda \|W\|_{\mathcal{F}}^2 \\ =&\min_W \|X - \mathcal{K} W Z\|_F^2 + \lambda \|W\|_{\mathcal{F}}^2 \end{aligned} \tag{9}$$

Note that (9) is a convex function of $W$ with a unique global minimizer.

Here we provide two solutions for this objective function.

**Solution 1: Iterative optimization**

We can rewrite (9) as

$$G(W) = \min_W \mathrm{trace}((X - \mathcal{K}WZ)^T (X - \mathcal{K}WZ)) + \lambda \mathrm{trace}(W^T W)$$

Taking the derivative with respect to W,

$$\nabla G(W) = \lambda W + KWZZ^T - Z\mathcal{K} = \lambda W + KWZZ^T - ZK^{\frac{1}{2}}$$

We find the optimality by using gradient descent.

**Solution 2: Find the explicit solution of (9) by vectorizing the matrices**

Vectorizing the matrices inside (9), the objective function is equivalent to

$$\begin{aligned} g(w) &= \min_{w=\text{vec}(W)} \left\| x - \frac{1}{n}\text{vec}(K^{\frac{1}{2}} W K^{\frac{1}{2}} X^{(1)}) \right\|_2^2 + \lambda \left\| w \right\|_2^2 \\ &= \min_{w=\text{vec}(W)} \left\| x - \frac{1}{n}(X^{(1)^T} K^{\frac{1}{2}}) \otimes K^{\frac{1}{2}} w \right\|_2^2 + \lambda \left\| w \right\|_2^2, \end{aligned} \tag{10}$$

where $x = \text{vec}(X), w = \text{vec}(W)$.

Denote $S = \frac{1}{n}(X^{(1)^T} K^{\frac{1}{2}}) \otimes K^{\frac{1}{2}}$, the objective function becomes

$$g(w) = \min_{w} \left\| x - Sw \right\|_2^2 + \lambda \left\| w \right\|_2^2 \tag{11}$$

We can obtain the explicit solution by solving the linear equation

$$(S^T S + \lambda I) w = S^T x \tag{12}$$

Notice that S is a $n^2 \times n^2$ matrix, solving (12) is very likely to be problematic. Here we further exploit the structure of $S^T S + \lambda I$. Denote $S_0 = K^{\frac{1}{2}} X^{(1)}/n$, we have $S^T S = S_0 S_0^T \otimes K$.

Do SVD to $S_0 S_0^T$ and $K$, we have $S_0 S_0^T = U_1 D_1 U_1^T$, $K = U_2 D_2 U_2^T$, $U_1 = (U_{11}, ..., U_{1n})$, $U_2 = (U_{21}, ..., U_{2n})$.

Hence,

$$\begin{aligned} &S^T S + \lambda I = S_0 S_0^T \otimes K + \lambda I = (U_1 \otimes U_2)(D_1 \otimes D_2 + \lambda I)(U_1^T \otimes U_2^T), \\ &(S^T S + \lambda I)^{-1} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\lambda + D_{1i} D_{2j}} (U_{1i} U_{1i}^T) \otimes (U_{2j} U_{2j}^T), \\ &\hat{w} = (S^T S + \lambda I)^{-1} S^T x = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\lambda + D_{1i} D_{2j}} [(U_{1i} U_{1i}^T) \otimes (U_{2j} U_{2j}^T)][S_0 \otimes K^{\frac{1}{2}}] x \end{aligned}$$

Therefore, the explicit expression of $\widehat{W}$ is

$$\widehat{W} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\lambda + D_{1i} D_{2j}} U_{2j} U_{2j}^T (K^{\frac{1}{2}} X X^{(1)^T} K^{\frac{1}{2}}) U_{1i} U_{1i}^T \tag{13}$$

Similar techniques can be found in Wang et al. (2020b).

Given $\widehat{W}$, the estimated transition operators $\{\widehat{A}\}$ can be recovered by

$$\widehat{A}(r, s) = k(r)^\top \widehat{R} k(s) = \frac{1}{\lambda} k(r)^\top K^{-\frac{1}{2}} \widehat{W} K^{-\frac{1}{2}} k(s) \tag{14}$$

For this solution, we use generalized cross validation to select the tuning parameter $\lambda$. The

GCV score is

$$V(\lambda) = \frac{\frac{1}{nT}\left\| X - \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\lambda+D_{1i}D_{2j}}(K^{\frac{1}{2}}U_{2j}U_{2j}^{T}K^{\frac{1}{2}})X(S_0^T U_{1i}U_{1i}^T S_0)\right\|_F^2}{(1-\frac{1}{nT}\mathrm{tr}(M(\lambda))^2}, \tag{15}$$

where

$$M(\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\lambda+D_{1i}D_{2j}}\mathrm{tr}(U_{1i}U_{1i}^{T}S_0S_0^T)\mathrm{tr}(U_{2j}U_{2j}^{T}K)$$

Plugging into (6), the one-step ahead prediction of $X_{T+1}$ is then

$$\widehat{X}_{T+1}(r) = \frac{1}{n}\frac{1}{\lambda}k(r)^{\top}K^{-\frac{1}{2}}\widehat{W}K^{\frac{1}{2}}X_T, \text{ for } r \in [0,1]. \tag{16}$$

# 3 One-dimensional Simulation Studies

In this section, we conduct simulation studies to study the estimation and prediction performance of the proposed penalized Frobenius and nuclear norm estimator and compare it with the state-of-the-art functional time series prediction method in Aue et al. (2015).

## 3.1 Basic simulation setting

**Data generating process**: We borrowed from the simulation setting in Aue et al. (2015). For $d = 1, 2, \cdots, D$, we assume the $d$th transition operator $A_d(r, s)$ is of rank $q_d$ and is generated by $q_d$ basis functions $\{u_i(s)\}_{i=1}^{q_d}$ such that

$$A_d(r, s) = \textstyle\sum_{i,j=1}^{q_d} \lambda_{d,ij}u_i(r)u_j(s),$$

where $\{u_i(s)\}_{i=1}^{q_d}$ consists of orthonormal basis of $\mathcal{L}^2[0, 1]$ that will be specified later. Define matrix $\Lambda_d$ such that $\Lambda_{d,ij} = \lambda_{d,ij}$ and define $\mathbf{u}_{q_d}(s) = (u_1(s), u_2(s), \cdots, u_{q_d}(s))^{\top}$. We have $A_d(r, s) = \mathbf{u}_{q_d}(r)^{\top}\Lambda_d\mathbf{u}_{q_d}(s)$. We further set the noise function $\epsilon_t$ to be of finite rank $q_\epsilon$ such that $\epsilon_t(s) = \sum_{i=1}^{q_\epsilon} z_{ti}u_i(s)$, where $z_{ti} \overset{i.i.d.}{\sim} U(-a_i, a_i)$ or $z_{ti} \overset{i.i.d.}{\sim} N(0, \sigma_i^2)$. Without loss of generality, we set $q_1 = q_2 = \cdots = q_D = q_\epsilon = q$ for simplicity. Hence, the FAR($D$) process $\{X_t(s)\}_{t=1}^{T}$ nests in a finite dimensional subspace spanned by the orthonormal basis $\{u_i(s)\}_{i=1}^{q}$. Denote $X_t(r) = \sum_{i=1}^{q} x_{ti}u_i(r)$ where $x_{ti} = \int X_t(r)u_i(r)dr$, and denote $x_t =$

$(x_{t1}, \cdots, x_{tq})^\top$ and $z_t = (z_{t1}, \cdots, z_{tq})^\top$. We have

$$\begin{aligned} X_t(r) &= \sum_{d=1}^{D} \int A_d(r,s) X_{t-d}(s) ds + \epsilon_t(r) = \sum_{d=1}^{D} \int \mathbf{u}_q(r)^\top \Lambda_d \mathbf{u}_q(s) X_{t-d}(s) ds + z_t^\top \mathbf{u}_q(r) \\ &= \sum_{d=1}^{D} \int \mathbf{u}_q(r)^\top \Lambda_d \mathbf{u}_q(s) \mathbf{u}_q(s)^\top x_{t-d} ds + z_t^\top \mathbf{u}_q(r) = \mathbf{u}_q(r)^\top \left( \sum_{d=1}^{D} \Lambda_d x_{t-d} + z_t \right). \end{aligned} \tag{17}$$

For FAR(1) process, the expression simplifies to

$$X_t(r) = \mathbf{u}_q(r)^\top \left( \Lambda x_{t-1} + z_t \right) \tag{18}$$

(40) leads to $x_t = \sum_{d=1}^{D} \Lambda_d x_{t-d} + z_t$. Thus, the FAR($D$) process can be exactly simulated via a VAR($D$) process. Following the simulation setting in Yuan and Cai (2010) and Sun et al. (2018), we set $u_i(s) = 1$ if $i = 1$ and $u_i(s) = \sqrt{2}\cos((i-1)\pi s)$ for $i = 2, \cdots, q$.

Given the transition operators $A_1, \cdots, A_D$ (i.e. $\Lambda_1, \cdots, \Lambda_D$) and the distribution of noise $z_t$, the true FAR($D$) process $\{X_t(s), s \in [0,1]\}_{t=1}^{T}$ can be simulated and discrete measurements of the functional time series are taken at the sampling points $\{s_i\}_{i=1}^{n}$. For simplicity, we set $\{s_i\}_{i=1}^{n}$ to be the $n$ equal-spaced points in $[0,1]$, which resembles the typical sampling scheme of functional time series in real data applications.

**Evaluation criteria**: We evaluate the performance of a method via (a). estimation error of $\widehat{A}_1, \widehat{A}_2, \cdots, \widehat{A}_D$ and (b). prediction error of the estimated FAR($D$) model.

Specifically, given sample size $(n, T)$, we simulate the observed functional time series $\{X_t(s_i), i = 1, \cdots, n\}_{t=1}^{T+0.2T}$, which we then partition into training data $\{X_t(s_i), i = 1, \cdots, n\}_{t=1}^{T}$ for estimation of $A_1, \cdots, A_D$ and test data $\{X_t(s_i), i = 1, \cdots, n\}_{t=T+1}^{T+0.2T}$ for evaluation of prediction performance. Denote $\{\widehat{X}_t(s_i), i = 1, \cdots, n\}_{t=T+1}^{T+0.2T}$ as the one-step ahead prediction given by the estimated FAR($D$) model. We define

$$\text{MISE}(\widehat{A}_d, A_d) = \int_{[0,1]} \int_{[0,1]} (A_d(r,s) - \widehat{A}_d(r,s))^2 dr ds \Big/ \int_{[0,1]} \int_{[0,1]} A_d(r,s)^2 dr ds, \tag{19}$$

$$\text{PE} = \frac{1}{0.2nT} \sum_{t=T+1}^{T+0.2T} \sum_{i=1}^{n} (X_t(s_i) - \widehat{X}_t(s_i))^2, \tag{20}$$

where MISE (mean integrated squared error) measures the estimation error and PE measures the prediction error.

## 3.2 Estimation methods and implementation details

We implement the functional PCA (FPCA) based estimation approach for FAR: the vector autoregressive based approach in Aue et al. (2015) for comparison. The estimator makes use of the FPCA conducted on the sample covariance operator $\widetilde{C}(s,r) = \frac{1}{T}\sum_{t=1}^{T} X_t(s)X_t(r)$ such that $\widetilde{C}(s,r) = \sum_{i=1}^{\infty} \hat{\lambda}_i \hat{f}_i(s)\hat{f}_i(r)$, where $(\hat{\lambda}_i, \hat{f}_i)$ is the eigenvalue-eigenfunction pair.

**Functional PCA-VAR estimator in Aue et al. (2015)** [ANH]: The basic idea of Aue et al. (2015) is the combination of FPCA-based dimension reduction and the classical vector autoregressive (VAR) model, designed for prediction of FAR processes. Specifically, the infinite dimensional functional time series $\{X_t\}_{t=1}^{T}$ is first projected to the $p$ eigenfunctions $\hat{f}(s) = (\hat{f}_1(s), \cdots, \hat{f}_p(s))^\top$ of the sample covariance operator. After projection, $X_t$ is represented by a $p$-dimensional functional principal score $x_t = (x_{t1}, \cdots, x_{tp})^\top$ with $x_{ti} = \int X_t(s)\hat{f}_i(s)ds$. A VAR($D$) model is then fitted on the $p$-dimensional time series $\{x_t\}_{t=1}^{T}$ such that $x_t = B_1 x_{t-1} + \cdots + B_D x_{t-D} + \epsilon_t$. Denote the estimated coefficient matrices as $\hat{B}_1, \cdots, \hat{B}_D \in \mathbb{R}^{p\times p}$, the one-step ahead prediction of $X_t(s)$ is then $\widehat{X}_t(s) = \hat{f}(s)^\top \hat{x}_t = \hat{f}(s)^\top \sum_{d=1}^{D} \hat{B}_d x_{t-d}$.

Note that this implies $\hat{X}_t(s) = \hat{f}(s)^\top \hat{x}_t = \hat{f}(s)^\top \sum_{d=1}^{D} \hat{B}_d x_{t-d} = \hat{f}(s)^\top \sum_{d=1}^{D} \hat{B}_d \int \hat{f}(r) X_{t-d}(r)dr = \sum_{d=1}^{D} \int \hat{f}(s)^\top \hat{B}_d \hat{f}(r) X_{t-d}(r)dr$. Thus, the FPCA-based prediction algorithm in Aue et al. (2015) induces an estimator for the transition operators $\{A_d\}_{d=1}^{D}$ such that

$$\widehat{A}_d(s,r) = \hat{f}(s)^\top \hat{B}_d \hat{f}(r), \text{ for } d = 1, \cdots, D.$$

We refer to this estimator by ANH.

**Implementation of FPCA-based estimators (ANH)**: For the implementation of ANH, the functional time series is required to be fully observed over the entire interval $[0,1]$. However, under the current simulation setting, only discrete measurements $\{X_t(s_i), i = 1, \cdots, n\}_{t=1}^{T}$ are available. Following Aue et al. (2015), for each $t$, the function $X_t(s), s \in [0,1]$ is estimated using 10 cubic B-spline basis functions based on the discrete measurements $(X_t(s_1), \cdots, X_t(s_n))$. We also use 20 cubic B-spline basis functions for more flexibility (see more details later).

**Implementation of penalized nuclear norm estimator (FAR-Tr)**: For the implementation of the proposed RKHS-based penalized nuclear norm estimator, we use the rescaled Bernoulli polynomial as the reproducing kernel $\mathbb{K}$, such that

$$\mathbb{K}(x,y) = 1 + k_1(x)k_1(y) + k_2(x)k_2(y) - k_4(x-y),$$

where $k_1(x) = x - 0.5$, $k_2(x) = \frac{1}{2}(k_1^2(x) - \frac{1}{12})$ and $k_4(x) = \frac{1}{24}(k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240})$ for $x \in [0,1]$, and $k_4(x-y) = k_4(|x-y|)$ for $x, y \in [0,1]$. Such $\mathbb{K}$ is the reproducing kernel for $W^{2,2}$. See Chapter 2.3.3 of Gu (2013) for more detail. The accelerated gradient algorithm in Ji and Ye (2009) is used to solve the trace norm minimization

**Implementation of penalized Frobenius norm estimator (FAR-F2)**: We implemented the two solutions provided in 2.2, and they give the same result. However, when tuning parameter $\lambda$ is relatively small, say $10^{-5}$, gradient descent becomes very slow to converge. Hence, in the following simulations, we use (13) to solve the optimization problem. Generalized cross validation is applied to select the tuning parameter $\lambda$. More details are in 2.2. A standard 5-fold cross validation is used to select the tuning parameter $\lambda$.

Based on $\{\widehat{A}\}$, the one-step ahead prediction of $X_t(s_i)$ for $t = T+1, \cdots, T+0.2T$ in the test data can be calculated via $\widehat{X}_t(s_i) = \frac{1}{n}\sum_{j=1}^{n} \widehat{A}(s_i, s_j) X_t(s_j)$ for $i = 1, \cdots, n$ as in (6).

## 3.3 Simulation result for FAR(1)

For FAR(1), there is only one transition operator $A(r,s) = A_1(r,s)$. The simulation setting involves the transition matrix $\Lambda = \Lambda_1 \in \mathbb{R}^{q \times q}$ (signal) and the noise range $a_{1:q} = (a_1, a_2, \cdots, a_q)$ or the noise variance $\sigma_{1:q}^2 = (\sigma_1^2, \sigma_2^2, \cdots, \sigma_q^2)$ for $\{z_{ti}\}_{i=1}^q$ (driving noise). Denote $\sigma(\Lambda)$ as the leading singular value for a matrix $\Lambda$. We consider two different signal-noise settings:

- Scenario A (Diag $\Lambda$): $\Lambda = \text{diag}(\kappa, \cdots, \kappa)$ and $z_{ti} \overset{i.i.d.}{\sim} U(-a, a)$ with $a = 0.1$ for $i = 1, \cdots, q$.
- Scenario B (Random $\Lambda$): A random matrix $\Lambda^*$ is first generated via $\Lambda_{ij}^* \overset{i.i.d.}{\sim} N(0,1)$ and we set $\Lambda = \kappa \cdot \Lambda^* / \sigma(\Lambda^*)$, and $z_{ti} \overset{i.i.d.}{\sim} U(-a, a)$ with $a = 0.1$ for $i = 1, \cdots, q$.

For ANH, the function $X_t(s)$ is first estimated using 10 cubic B-spline basis functions. We use 20 cubic B-splines when $q = 21$ for more flexibility. With the FAR order fixed at

$D = 1$, the fFPE criterion is used to select the number of FPCs $p$ for ANH. For RKHS, we implemented the two solutions provided in 2.2, and they gives the same result. However, when tuning parameter $\lambda$ is relatively small, say $10^{-5}$, gradient descent becomes very slow to converge. Hence, in the following simulations, we use (13) to solve the optimization problem. Generalized cross validation is applied to select the tuning parameter $\lambda$. More details are in 2.2.

**Numerical result for FAR(1)**: For Scenarios A and B, we consider three sample sizes: $(1) q = 6, n = 20, T = 100$, $(2) q = 12, n = 20, T = 400$, $(3) q = 21, n = 40, T = 400$. As for the signal level, we vary the spectral norm of $\Lambda$ by $\kappa = 0.2, 0.5, 0.8$. For each simulation setting, i.e. different combination of Scenario A,B and $(q, n, T, \kappa)$, we conduct 20 experiments. Note that the transition matrix $\Lambda$ is randomly generated for each experiment under Scenario B.

We summarize the numerical performance of ANH, FAR-F2 and FAR-Tr in Table 2, where we report the mean MISE (MISE$_{avg}$) and mean PE (PE$_{avg}$) across the 20 experiments. For each experiment, we also calculate the percentage improvement of prediction by FAR-F2 or FAR-Tr over ANH via Ratio= (PE(ANH) / PE(FAR-x)$-1$) $\times$ 100%. A positive ratio indicates improvement by FAR-F2 or FAR-Tr. We report the mean ratio (denoted by R$_{avg}$) across the 20 experiments.

Overall, FAR-F2, FAR-Tr and ANH are suitable for different situations. Within each scenario, the improvement of RKHS based methods over comparison methods increase with a higher dimension $q$ and a stronger signal $\kappa$, while for the same $(q, \kappa)$, the 2 RKHS based method yield the most improvement in Scenario A when spectral norm $\kappa$ and effective dimension $q$ are relatively large.

|  |  | Scenario A: $q = 6, n = 20, T = 100$ |  |  | Scenario B: $q = 6, n = 20, T = 100$ |  |  |
|---|---|---|---|---|---|---|---|
|  | Method | MISE$_{avg}$ | PE$_{avg}\times$100 | R$_{avg}$(%) | MISE$_{avg}$ | PE$_{avg}\times$100 | R$_{avg}$(%) |
| $\kappa = 0.2$ | FAR-F2 | 1.373 | 2.198 | -2.61 | 2.066 | 2.228 | -4.85 |
|  | ANH | **0.898** | **2.141** |  | **1.958** | **2.120** |  |
| $\kappa = 0.5$ | FAR-F2 | **0.274** | **2.231** | 1.82 | **0.587** | **2.226** | 0.00 |
|  | ANH | 0.316 | 2.271 |  | 0.742 | 2.226 |  |
| $\kappa = 0.8$ | FAR-F2 | 0.106 | 2.350 | -1.50 | **0.307** | **2.206** | 1.38 |
|  | ANH | **0.066** | **2.315** |  | 0.349 | 2.237 |  |
|  |  | Scenario A: $q = 12, n = 20, T = 400$ |  |  | Scenario B: $q = 12, n = 20, T = 400$ |  |  |
|  | Method | MISE$_{avg}$ | PE$_{avg}\times$100 | R$_{avg}$(%) | MISE$_{avg}$ | PE$_{avg}\times$100 | R$_{avg}$(%) |
| $\kappa = 0.2$ | FAR-F2 | 0.809 | 4.299 | -0.53 | 1.368 | 4.224 | -0.28 |
|  | ANH | **0.715** | **4.276** |  | **1.112** | **4.212** |  |
| $\kappa = 0.5$ | FAR-F2 | **0.150** | **4.365** | 5.63 | **0.416** | **4.327** | 1.62 |
|  | ANH | 0.417 | 4.611 |  | 1.088 | 4.400 |  |
| $\kappa = 0.8$ | FAR-F2 | **0.078** | **4.365** | 38.94 | **0.154** | **4.306** | 8.42 |
|  | ANH | 0.343 | 6.065 |  | 0.923 | 4.668 |  |
|  |  | Scenario A: $q = 21, n = 40, T = 400$ |  |  | Scenario B: $q = 21, n = 40, T = 400$ |  |  |
|  | Method | MISE$_{avg}$ | PE$_{avg}\times$100 | R$_{avg}$(%) | MISE$_{avg}$ | PE$_{avg}\times$100 | R$_{avg}$(%) |
| $\kappa = 0.2$ | FAR-F2 | 0.976 | 7.534 | -1.59 | **1.009** | 7.386 | -1.23 |
|  | ANH | **0.833** | **7.414** |  | 1.065 | **7.259** |  |
| $\kappa = 0.5$ | FAR-F2 | **0.231** | **7.769** | 2.83 | **0.653** | **7.512** | 1.18 |
|  | ANH | 0.340 | 7.988 |  | 3.342 | 7.601 |  |
| $\kappa = 0.8$ | FAR-F2 | **0.080** | **7.794** | 19.52 | **0.268** | **7.591** | 4.39 |
|  | ANH | 0.308 | 9.315 |  | 3.698 | 7.925 |  |

Table 1: Numerical performance of various methods for FAR(1) processes. Methods considered are RKHS (this paper), ANH (Aue et al., 2015). Bold font indicates the best performance, where the proposed RKHS method is generally the best performer in Scenarios A and B with large spectral norm $\kappa$.

| | | Scenario A: $q = 6, n = 20, T = 100$ | | | Scenario B: $q = 6, n = 20, T = 100$ | | |
|---|---|---|---|---|---|---|---|
| | Method | $\text{MISE}_{avg}$ | $\text{PE}_{avg} \times 100$ | $\text{R}_{avg}(\%)$ | $\text{MISE}_{avg}$ | $\text{PE}_{avg} \times 100$ | $\text{R}_{avg}(\%)$ |
| $\kappa = 0.2$ | FAR-F2 | 1.386 | 2.191 | -1.69 | 2.077 | 2.166 | -1.06 |
| | FAR-Tr | 0.970 | **2.144** | 0.47 | **1.529** | **2.138** | 0.23 |
| | ANH | **0.882** | 2.154 | | 1.532 | 2.143 | |
| $\kappa = 0.5$ | FAR-F2 | 0.278 | 2.138 | 3.32 | **0.741** | 2.163 | 1.80 |
| | FAR-Tr | **0.232** | **2.137** | 3.32 | **0.741** | **2.160** | 1.94 |
| | ANH | 0.312 | 2.209 | | 0.845 | 2.202 | |
| $\kappa = 0.8$ | FAR-F2 | 0.113 | 2.399 | 0.79 | 0.222 | **2.265** | 0.08 |
| | FAR-Tr | **0.083** | **2.392** | 1.09 | **0.195** | 2.266 | 0.04 |
| | ANH | 0.086 | 2.418 | | 0.296 | 2.267 | |
| | | Scenario A: $q = 12, n = 20, T = 400$ | | | Scenario B: $q = 12, n = 20, T = 400$ | | |
| | Method | $\text{MISE}_{avg}$ | $\text{PE}_{avg} \times 100$ | $\text{R}_{avg}(\%)$ | $\text{MISE}_{avg}$ | $\text{PE}_{avg} \times 100$ | $\text{R}_{avg}(\%)$ |
| $\kappa = 0.2$ | FAR-F2 | 0.871 | **4.249** | 0.87 | 1.388 | 4.173 | 0.02 |
| | FAR-Tr | **0.635** | 4.260 | 0.61 | **1.025** | **4.169** | 0.12 |
| | ANH | 0.715 | 4.286 | | 1.062 | 4.174 | |
| $\kappa = 0.5$ | FAR-F2 | 0.152 | 4.495 | 6.85 | 0.388 | 4.412 | 2.79 |
| | FAR-Tr | **0.137** | **4.458** | 7.74 | **0.329** | **4.380** | 3.54 |
| | ANH | 0.454 | 4.803 | | 1.341 | 4.535 | |
| $\kappa = 0.8$ | FAR-F2 | 0.081 | 4.287 | 33.68 | **0.154** | **4.122** | 15.14 |
| | FAR-Tr | **0.079** | **4.221** | 35.77 | 0.166 | 4.129 | 14.94 |
| | ANH | 0.338 | 5.731 | | 0.737 | 4.746 | |

Table 2: Numerical performance of various methods for FAR(1) processes. Bold font indicates the best performance.

Notice that although FAR-Tr is better than FAR-F2 in some cases, the nuclear norm penalty is much slower than FAR-F2, which might cause problem when it comes to multi-dimensional setting.

# 4 Two-dimensional Scenario Extension

## 4.1 Two-dimensional RKHS-based Penalized Estimator

We construct its estimator $\{\widehat{A}_d\}_{d=1}^D$ via a constrained Frobenius norm optimization such that

$$\{\widehat{A}_d\}_{d=1}^D = \underset{\{A_d\}_{d=1}^D \in \mathcal{C}_\tau}{\arg\min} \frac{1}{Tn} \sum_{t=D+1}^{T} \sum_{i=1}^{n} \left( X_t(s_i, r_i) - \sum_{d=1}^{D} \frac{1}{n} \sum_{j=1}^{n} A_d((s_i, r_i), (s_j, r_j)) X_{t-d}(s_j, r_j) \right)^2 \tag{21}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_D)$ is the tuning parameter and $\mathcal{C}_{\boldsymbol{\tau}} := \{(A_1, \cdots, A_D) : A_d \in \mathcal{C}$ and $\|A_d\|_{\mathcal{H},\mathcal{F}} \leq \tau_d,\ d = 1, \cdots, D\}$ is the constraint space. We name $\{\widehat{A}_d\}_{d=1}^D$ in (21) the penalized/constrained Frobenius norm estimator for transition operators of FAR. Similarly, we extend the representer theorem into 2 dimensional case.

**Proposition 2** (Representer theorem)**.** *There exists a minimizer $\{\widehat{A}_d\}_{d=1}^D$ of the constrained nuclear/Frobenius norm optimization* (21) *such that for any $(r, s) = ((r^1, r^2), (s^1, s^2)) \in [0, 1] \times [0, 1] \times [0, 1] \times [0, 1]$,*

$$\widehat{A}_d(r, s) = \widehat{A}_d((r^1, r^2), (s^1, s^2)) = \sum_{1 \leq i,j \leq n} \widehat{a}_{d,ij} \mathbb{K}((r^1, r^2), (s_i^1, s_i^2)) \mathbb{K}((s^1, s^2), (s_j^1, s_j^2)), \tag{22}$$

*for $d = 1, 2, \cdots, D$.*

Plugging the representer theorem into (21), the optimization problem is almost the same as in 1 dimensional scenario, except that the input is spatial correlated. Hence, the extension to 2 dimensional Frobenius norm penalty estimator is trivial(See Algorithm 1). In the following section, we will introduce the 2 dimensional penalized functional PCA in detail, which will serve as a baseline method.

**Algorithm 1** FAR-F2

```
INPUT:$X, X^{(1)}$, n sample points $(s_1^1, s_1^2), ..., (s_n^1, s_n^2)$, kernel matrix $K$.
Do SVD to $\frac{1}{n^2} K^{\frac{1}{2}} X^{(1)} {X^{(1)}}^T K^{\frac{1}{2}} = U_1 D_1 U_1^T, K = U_2 D_2 U_2^T$.
Plug into 13 ▷ $U_{ij}$ is the $j_{\text{th}}$ column of matrix $U_i$, $D_{ij}$ is the $j_{\text{th}}$ element of the diagonal of matrix $D_i$.
OUTPUT:$\hat{W}, \hat{A}, \hat{X}$ by 14 and 16
```

## 4.2 Penalized Functional PCA

The framework for 1 dimensional functional PCA is described in Aue et al. (2015). The basic idea of Aue et al. (2015) is a canny combination of FPCA-based dimension reduction and the classical vector autoregressive (VAR) model, designed for prediction of FAR processes. Specifically, the infinite dimensional functional time series $\{X_t\}_{t=1}^T$ is first projected to the $p$ eigenfunctions $\hat{f}(s) = (\hat{f}_1(s), \cdots, \hat{f}_p(s))^\top$ of the sample covariance operator. After projection, $X_t$ is represented by a $p$-dimensional functional principal score

$x_t = (x_{t1}, \cdots, x_{tp})^\top$ with $x_{ti} = \int X_t(s)\hat{f}_i(s)ds$. A VAR($D$) model is then fitted on the $p$-dimensional time series $\{x_t\}_{t=1}^T$ such that $x_t = B_1 x_{t-1} + \cdots + B_D x_{t-D} + \epsilon_t$. Denote the estimated coefficient matrices as $\hat{B}_1, \cdots, \hat{B}_D \in \mathbb{R}^{p\times p}$, the one-step ahead prediction of $X_t(s)$ is then $\widehat{X}_t(s) = \hat{f}(s)^\top \hat{x}_t = \hat{f}(s)^\top \sum_{d=1}^D \hat{B}_d x_{t-d}$.

Based on the 1 dimension framework, the only difference for 2 dimension algorithm is to obtain the eigenfunctions with spatial data input. To be more specific, the only thing we need to work on is to extend functional PCA into 2 dimensional case. The one dimensional fPCA method has been stated in Ramsay and Silverman (2005). In this section, we will extend functional PCA to 2 dimension scenario together with smoothing spline method by using the tensor products of B-splines.

Given sample points $(s_1, r_1)$, ..., $(s_n, r_n)$, the response functional data $X_t(s,r)$ where $(s,r) \in \mathbb{R}^2$ and the spline basis $\Phi_1(s) = (\phi_{11}(s), ..., \phi_{1k_1}(s))^T$, $\Phi_2(r) = (\phi_{21}(r), ..., \phi_{2k_2}(r))^T$ for 2 dimensions respectively, we define the 2 dimensional basis function as

$$\begin{aligned} \Phi(s,r) &= (\phi_{11}(s)\phi_{21}(r), ..., \phi_{1k_1}(s)\phi_{21}(r), \phi_{11}(s)\phi_{22}(r), ...)^T \\ &= \Phi_2(r) \otimes \Phi_1(s) \\ &= (\Phi_2(r) \otimes \mathbb{1}_{k_1}(s)) \odot (\mathbb{1}_{k_2}(r) \otimes \Phi_1(s)) \end{aligned} \tag{23}$$

where $\odot$ is the Hadamard product. (23) was proposed by Lee and Durbán (2011) in multi-dimensional smoothing.

Therefore, we have $X_t(s,r) = c_t^T \Phi(s,r)$. Let $y$ be the vector of coefficients of any potential principle component curve $\xi(s,r)$, so that $\xi(s,r) = y^T \Phi(s,r)$. As stated in Ramsay and Silverman (2005), we need to minimize the penalized sample variance, which is defined as

$$\text{PCAPSV}(\xi) = \frac{\textbf{var} \iint \xi X_t}{||\xi||^2 + \lambda_1 \times \text{PEN}_2(\xi)} \tag{24}$$

In our paper, we define

$$\begin{aligned}
\mathrm{PEN}_2(\xi) &\coloneqq ||D^2\xi||_F^2 = \iint (\frac{\partial^2\xi}{\partial s^2})^2 + 2(\frac{\partial^2\xi}{\partial s\partial r})^2 + (\frac{\partial^2\xi}{\partial r^2})^2 \\
&= y^T \left[\iint \left(\frac{\partial^2\Phi}{\partial s^2}\left(\frac{\partial^2\Phi}{\partial s^2}\right)^T + 2\frac{\partial^2\Phi}{\partial s\partial r}\left(\frac{\partial^2\Phi}{\partial s\partial r}\right)^T + \frac{\partial^2\Phi}{\partial r^2}\left(\frac{\partial^2\Phi}{\partial r^2}\right)^T\right)\right] y \\
&\coloneqq y^T(K_1 + 2K_2 + K_3)y \\
&\coloneqq y^T K y. \qquad (25)
\end{aligned}$$

From (23) we have,

$$\begin{aligned}
\frac{\partial^2\Phi}{\partial s^2} &= (\Phi_2(r) \otimes \mathbb{1}_{k_1}(s)) \odot \left(\mathbb{1}_{k_2}(r) \otimes D^2\Phi_1(s)\right) \\
&= (\Phi_2(r) \odot \mathbb{1}_{k_2}(r)) \otimes \left(\mathbb{1}_{k_1}(s) \odot D^2\Phi_1(s)\right) \\
&= \Phi_2(r) \otimes D^2\Phi_1(s) \qquad (26)
\end{aligned}$$

Similarly,

$$\frac{\partial^2\Phi}{\partial s\partial r} = (D\Phi_2(r) \otimes \mathbb{1}_{k_1}(s)) \odot (\mathbb{1}_{k_2}(r) \otimes D\Phi_1(s)) = D\Phi_2(r) \otimes D\Phi_1(s) \qquad (27)$$

$$\frac{\partial^2\Phi}{\partial r^2} = \left(D^2\Phi_2(r) \otimes \mathbb{1}_{k_1}(s)\right) \odot (\mathbb{1}_{k_2}(r) \otimes \Phi_1(s)) = D^2\Phi_2(r) \otimes \Phi_1(s) \qquad (28)$$

Therefore,

$$\begin{aligned}
K_1 &= \iint \frac{\partial^2\Phi}{\partial s^2}\left(\frac{\partial^2\Phi}{\partial s^2}\right)^T \\
&= \iint \left(\Phi_2(r) \otimes D^2\Phi_1(s)\right) \cdot \left(\Phi_2^T(r) \otimes D^2\Phi_1^T(s)\right) \\
&= \iint \left(\Phi_2(r)\Phi_2^T(r)\right) \otimes \left(D^2\Phi_1(s)D^2\Phi_1^T(s)\right) \\
&= \int \Phi_2(r)\Phi_2^T(r) \otimes \int D^2\Phi_1(s)D^2\Phi_1^T(s) \qquad (29)
\end{aligned}$$

Likewise,

$$\begin{aligned}
K_2 &= \int D\Phi_2(r)D\Phi_2^T(r) \otimes \int D\Phi_1(s)D\Phi_1^T(s) \\
K_3 &= \int D^2\Phi_2(r)D^2\Phi_2^T(r) \otimes \int \Phi_1(s)\Phi_1^T(s) \qquad (30)
\end{aligned}$$

We then calculate other terms in (24)

$$\textbf{var} \iint \xi X_t = \textbf{var}\ y^T \iint \Phi(s,r)\Phi(s,r)^T c_t := y^T J V J^T y \tag{31}$$

$$||\xi||^2 = \iint \xi^2(s,r) ds dr = y^T J y \tag{32}$$

where $V = \textbf{var}\ c_t$, $J = \iint \Phi(s,r)\Phi(s,r)^T$. Let $C = (c_1, ..., c_T)$, $V = CC^T/T$.

To further exploit the structure of $J$, we have

$$J = \iint \Phi(s,r)\Phi(s,r)^T = \int \Phi_2(r)\Phi_2^T(r) \otimes \int \Phi_2(s)\Phi_2^T(s) := J_2 \otimes J_1 \tag{33}$$

Hence, $K = J_2 \otimes D^2 J_1 + 2DJ_2 \otimes DJ_1 + D^2 J_2 \otimes J_1$.

Plug (25), (31), (29) and (30) into (24), the minimization problem becomes the eigenequation

$$JVJy = \rho(J + \lambda_1 K)y \tag{34}$$

$J$ and $K$ can be expressed as the tensor product of the corresponding matrices in 1 dimensional case. In the following context, we will find the expression of coefficient matrix $C$.

Denote the observation of the functional time series at time $t$ as $X_t = (X_t(s_1,r_1), X_t(s_2,r_2), \cdots, X_t(s_n,r_n))^\top$. Data $X_t$s are assumed to be the result of the model

$$X_t = \widetilde{\Phi} c_t + \epsilon_t \tag{35}$$

$\widetilde{\Phi}$ is the matrix of B-splines,

$$\widetilde{\Phi} = (\Phi(s_1,r_1), .., \Phi(s_n,r_n))^T = (\Phi_2(r_1) \otimes \Phi_1(s_1), ..., \Phi_2(r_n) \otimes \Phi_1(s_n))^T \tag{36}$$

Define $X = (X_1, ..., X_T)$, we have

$$X = \widetilde{\Phi} C + \epsilon \tag{37}$$

where $C = (c_1, ..., c_T)$.

To estimate $\theta$, the penalized B-spline is here considered, minimizing the penalized optimization function

$$f = (X - \widetilde{\Phi} C)^T (X - \widetilde{\Phi} C) + \lambda_2 C^T K C \tag{38}$$

The solution of minimization problem (38) is

$$C = (\widetilde{\Phi}^T\widetilde{\Phi} + \lambda_2 K)^{-1}\widetilde{\Phi}^T X \tag{39}$$

Plug (39) into (34), where $V = \textbf{var}\ c_t = CC^T/T$. Solve (34) and we will have coefficients of the eigenfunction.

The explicit algorithm is summarized in Algorithm 2. We coded it based on 1 dimensional functional PCA method, and took advantage of the R package 'fda'. Hence, we could get basis $\Phi(s,r), \widetilde{\Phi}$ by combining 1 dimensional b-spline basis with 23 and 36. $J_1 = \int \Phi_1(s)\Phi_1(s)^T, J_2 = \int \Phi_2(s)\Phi_2(s)^T, DJ_1, DJ_2, D^2J_1, D^2J_2$ defined in 26 - 33 can be obtained using functions in **fda** package.

**Algorithm 2** 2D FPCA

```
INPUT:$X \in \mathbb{R}^{n\times T}$, n sample points $(s_1^1, s_1^2), ..., (s_n^1, s_n^2)$, $\lambda_1$, $\lambda_2$.
Obtain $J_1, J_2, \widetilde{\Phi}, \Phi$ using package fda.
$K = J_2 \otimes D^2J_1 + 2DJ_2 \otimes DJ_1 + D^2J_2 \otimes J_1$ and plug into 39 to get $\hat{C}$.
Solve 34, where $V = \hat{C}\hat{C}^T/T$
OUTPUT:y is the coefficient of the eigenfunction.
```

Algorithm 2 outputs the 2 dimensional eigenfunction. Combining with Algorithm 1 in Aue et al. (2015), the method could be used in 2 dimensional functional prediction problem.

# 5 Two-dimensional Simulation Studies

In this section, we conduct simulation studies to investigate the estimation and prediction performance of the 2 dimensional penalized Frobenius norm estimator and compare it with the 2 dimensional functional PCA method in section 3.

## 5.1 Estimation methods and implementation details

**Data generating process**: Similar to 1 dimensional data generation process, we first define an FAR($D$) process, borrowed from the simulation setting in Aue et al. (2015), that is used in the simulation study. For $d = 1, 2, \cdots, D$, we assume the $d$th transition operator $A_d(r, s)$

is of rank $q_d$ and is generated by $q_d$ basis functions $\{u_i(s)\}_{i=1}^{q_d}$ such that

$$A_d((s,r),(l,w)) = \textstyle\sum_{i,j=1}^{q_d} \lambda_{d,ij} u_i(s,r) u_j(l,w),$$

where $\{u_i(s)\}_{i=1}^{q_d}$ consists of orthonormal basis of $\mathcal{L}^2[0,1]$ that will be specified later. Define matrix $\Lambda_d$ such that $\Lambda_{d,ij} = \lambda_{d,ij}$ and define $\mathbf{u}_{q_d}(s) = (u_1(s,r), u_2(s,r), \cdots, u_{q_d}(s,r))^\top$. We have $A_d(r,s) = \mathbf{u}_{q_d}(r)^\top \Lambda_d \mathbf{u}_{q_d}(s)$. We further set the noise function $\epsilon_t$ to be of finite rank $q_\epsilon$ such that $\epsilon_t(s) = \sum_{i=1}^{q_\epsilon} z_{ti} u_i(s,r)$, where $z_{ti} \overset{i.i.d.}{\sim} U(-a_i, a_i)$ or $z_{ti} \overset{i.i.d.}{\sim} N(0, \sigma_i^2)$.

Without loss of generality, we set $q_1 = q_2 = \cdots = q_D = q_\epsilon = q$ for simplicity. Thus, the FAR($D$) process $\{X_t(s)\}_{t=1}^T$ resides in a finite dimensional subspace spanned by the orthonormal basis $\{u_i(s)\}_{i=1}^q$. Denote $X_t(r) = \sum_{i=1}^q x_{ti} u_i(r)$ where $x_{ti} = \int X_t(r) u_i(r) dr$, and denote $x_t = (x_{t1}, \cdots, x_{tq})^\top$ and $z_t = (z_{t1}, \cdots, z_{tq})^\top$. We have

$$\begin{aligned}
X_t(s,r) =& \sum_{d=1}^D \iint A_d((s,r),(l,w)) X_{t-d}(l,w) dl dw + \epsilon_t(s,r) \\
=& \sum_{d=1}^D \iint \mathbf{u}_q(s,r)^\top \Lambda_d \mathbf{u}_q(l,w) X_{t-d}(l,w) dl dw + z_t^\top \mathbf{u}_q(s,r) \\
=& \sum_{d=1}^D \iint \mathbf{u}_q(s,r)^\top \Lambda_d \mathbf{u}_q(l,w) \mathbf{u}_q(l,w)^\top x_{t-d} dl dw + z_t^\top \mathbf{u}_q(s,r) \\
=& \ \mathbf{u}_q(s,r)^\top \left( \sum_{d=1}^D \Lambda_d x_{t-d} + z_t \right). \qquad (40)
\end{aligned}$$

For FAR(1) process, the expression simplifies to

$$X_t(s,r) = \mathbf{u}_q(s,r)^\top \left( \sum_{d=1}^D \Lambda_d x_{t-d} + z_t \right) \qquad (41)$$

(40) leads to $x_t = \sum_{d=1}^D \Lambda_d x_{t-d} + z_t$. Thus, the FAR($D$) process can be exactly simulated via a VAR($D$) process. We set $u_i(s,r) = 1$ if $i = 1$ and $u_i(s,r) = \sqrt{2}\cos(2(i-1)\pi(s+r))$ for $i = 2, \cdots, q$.

Given the transition operators $A_1, \cdots, A_D$ (i.e. $\Lambda_1, \cdots, \Lambda_D$) and the distribution of noise $z_t$, the true FAR($D$) process $\{X_t(s,r), (s,r) \in [0,1] \times [0,1]\}_{t=1}^T$ can be simulated and discrete measurements of the functional time series are taken at the sampling points $\{(s_i, r_i)\}_{i=1}^n$. For simplicity, we set $\{(s_i, r_i)\}_{i=1}^n$ to be the $n$ equal-spaced points on $[0,1] \times [0,1]$, which resembles the typical sampling scheme of functional time series in real data applications.

**Implementation of penalized Functional PCA-VAR estimator(ANH)**: There are 2 tuning parameters in the 2 dimensional Functional PCA-VAR method: $\lambda_1$ from the penalized functional PCA, $\lambda_2$ is from smoothing spline. In the following experiments, we simplify the case into $\lambda_1 = \lambda_2 = \lambda$. We select the number of functional principal component $p$ by looking at the proportion of variance explained. Further work on dimension selection could be exploit in the future.

**Implementation of penalized Frobenius norm estimator(FAR-F2)**: For the implementation of the proposed RKHS-based penalized Frobenius norm estimator, we use the Gaussian function as the reproducing kernel $\mathbb{K}$, such that

$$\mathbb{K}(\mathbf{x}, \mathbf{y}) = \frac{e^{||\mathbf{x}-\mathbf{y}||^2}}{2\sigma^2}$$

Generalized cross validation is applied to select the tuning parameter $\lambda$. More details are in 2.2.

**Implementation of penalized nuclear norm estimator(FAR-Tr)**: We use the accelerated gradient algorithm in Ji and Ye (2009) to solve the nuclear norm optimization problem. The standard 5 fold cross validation is used to tune the parameter $\lambda$. The reproducing kernel is the same as in method FAR-F2.

Based on $\{\widehat{A}\}$, the one-step ahead prediction of $X_t(s_i, r_i)$ for $t = T+1, \cdots, T+0.2T$ in the test data can be calculated via $\widehat{X}_t(s_i) = \frac{1}{n}\sum_{j=1}^{n} \widehat{A}((s_i, r_i), (s_j, r_j))X_t(s_j, r_j)$ for $i = 1, \cdots, n$ as in (6).

**Evaluation criteria**: We evaluate the performance of a method via (a). estimation error of $\widehat{A}_1, \widehat{A}_2, \cdots, \widehat{A}_D$ and (b). prediction error of the estimated FAR($D$) model.

Specifically, given sample size $(n, T)$, we simulate the observed functional time series $\{X_t(s_i, r_i), i = 1, \cdots, n\}_{t=1}^{T+0.2T}$, which we then partition into training data $\{X_t(s_i, r_i), i = 1, \cdots, n\}_{t=1}^{T}$ for estimation of $A_1, \cdots, A_D$ and test data $\{X_t(s_i, r_i), i = 1, \cdots, n\}_{t=T+1}^{T+0.2T}$ for evaluation of prediction performance. Denote $\{\widehat{X}_t(s_i), i = 1, \cdots, n\}_{t=T+1}^{T+0.2T}$ as the one-step ahead prediction given by the estimated FAR($D$) model. We define

$$\text{MISE}(\widehat{A}_d, A_d) = \frac{\iint_{[0,1]\times[0,1]} \iint_{[0,1]\times[0,1]} (A_d((s,r),(l,w)) - \widehat{A}_d((s,r),(l,w)))^2 dsdrdldw}{\iint_{[0,1]\times[0,1]} \iint_{[0,1]\times[0,1]} A_d((s,r),(l,w))^2 dsdrdldw} \tag{42}$$

$$\text{PE} = \frac{1}{0.2nT} \sum_{t=T+1}^{T+0.2T} \sum_{i=1}^{n} (X_t(s_i, r_i) - \widehat{X}_t(s_i, r_i))^2, \tag{43}$$

where MISE (mean integrated squared error) measures the estimation error and PE measures the prediction error.

## 5.2 Simulation result for FAR(1)

For FAR(1), there is only one transition operator $A((s,r),(l,w)) = A_1((s,r),(l,w))$. The simulation setting involves the transition matrix $\Lambda = \Lambda_1 \in \mathbb{R}^{q\times q}$ (signal) and the noise range $a_{1:q} = (a_1, a_2, \cdots, a_q)$ or the noise variance $\sigma^2_{1:q} = (\sigma^2_1, \sigma^2_2, \cdots, \sigma^2_q)$ for $\{z_{ti}\}_{i=1}^q$ (driving noise). Denote $\sigma(\Lambda)$ as the leading singular value for a matrix $\Lambda$. We consider two different signal-noise settings:

- Scenario A (Diag $\Lambda$): $\Lambda = \text{diag}(\kappa, \cdots, \kappa)$ and $z_{ti} \overset{i.i.d.}{\sim} U(-a, a)$ with $a = 0.1$ for $i = 1, \cdots, q.$, $\sigma$ of kernel function = 0.1. The function $X_t(s)$ is first estimated using 9*10 cubic penalized B-spline basis functions. We let sample size $n = 100$, effective dimension $q = 6, 6^2$.
- Scenario C (Low-rank setting): Partition the spatial domain $[0,1]\times[0,1]$ into 4 regions $I, II, III, IV$.

$$A((s,r),(s',r')) = \begin{cases} \kappa & (s,r),(s',r') \text{ are in the same region} \\ 0 & \text{Otherwise} \end{cases}$$

In scenario C, Let FAR(1) process $\{X_t(s)\}_{t=1}^T$ resides in a finite dimensional subspace spanned by the orthonormal basis $\{u_i(s)\}_{i=1}^q$. Denote $X_t(s,r) = \sum_{i=1}^q x_{ti} u_i(s,r)$ where $x_{ti} = \iint X_t(s,r) u_i(s,r) dsdr$, and denote $x_t = (x_{t1}, \cdots, x_{tq})^\top$ and $z_t = (z_{t1}, \cdots, z_{tq})^\top$, we will have $X_t(s,r) = x_t^T u(s,r), x_t = \iint X_t(s,r) u_(s,r) dsdr$. Assume $(s,r) \in$ region R, we have

$$\begin{aligned}
X_t(s,r) =& \iint A((s,r),(s',r'))X_{t-1}(s',r')ds'dr' + \epsilon_t(s,r) \\
=& \iint_R X_{t-1}(s',r')ds'dr' + \epsilon_t(s,r) \\
=& x_{t-1}^T \iint_R u(s',r')ds'dr' + \epsilon_t(s,r) \\
=& x_{t-1}^T \widetilde{u}(s,r) + \epsilon_t(s,r)
\end{aligned} \tag{44}$$

where piecewise function $\widetilde{u}(s,r) = \iint_R u(s',r')ds'dr'$, if $(s,r) \in$ region R, for $R = I, II, III, IV$. Denote $\widetilde{u}(R) = \widetilde{u}(s,r)$, if $(s,r) \in$ region R,

$$\begin{aligned}
x_t =& \iint X_t(s,r)u_(s,r)dsdr \\
=& \iint (x_{t-1}^T \widetilde{u}(s,r))u(s,r)dsdr + \iint \epsilon_t(s,r)u_(s,r)dsdr \\
=& \sum_{R=I,II,III,IV} (\iint_R u(s,r)\widetilde{u}(s,r)^T dsdr)x_{t-1} + z_t \\
=& \sum_{R=I,II,III,IV} (\iint_R u(s,r)dsdr)\widetilde{u}(R)^T x_{t-1} + z_t \\
=& \sum_{R=I,II,III,IV} \widetilde{u}(R)\widetilde{u}(R)^T x_{t-1} + z_t \\
:=& \Lambda x_{t-1} + z_t
\end{aligned} \tag{45}$$

where $\Lambda = \sum_{R=I,II,III,IV} \widetilde{u}(R)\widetilde{u}(R)^T$, $\widetilde{u}(R) = \iint_R u(s',r')ds'dr'$.

In this setting, we let region $I = [0, \frac{1}{2}] \times [\frac{1}{2}, 1]$, $II = [\frac{1}{2}, 1] \times [\frac{1}{2}, 1]$, $III = [0, \frac{1}{2}] \times [0, \frac{1}{2}]$, $IV = [\frac{1}{2}, 1] \times [0, \frac{1}{2}]$. Let $u_i(s,r) = 1$ if $i = 1$ and $u_i(s,r) = \sqrt{2}\cos(2(i-1)\pi(s+r))$ for $i = 2, \cdots, q$,

$$\widetilde{u}_i(R) = \begin{cases} \frac{1}{4} & i = 1 \\ \sqrt{2}\frac{S_i(R)}{\lambda_i^2} & i = 2, ..., q \end{cases} \tag{46}$$

where denote the range for x axis of region R is from $w_1$ to $w_2$, while the range for y axis of region R is from $z_1$ to $z_2$, $\lambda_i = 2(i-1)\pi$, $S_i(R) = (sin(kw_2) - sin(kw_1))(sin(kz_2) - sin(kz_1)) - (cos(kw_2) - cos(kw_1))(cos(kz_2) - cos(kz_1))$.

**Numerical result for FAR(1)**: For Scenarios A and C, we consider four sample sizes: (1)$q = 6, n = 36, T = 100$, (2)$q = 6, n = 100, T = 400$, (3)$q = 6^2, n = 36, T = 400$, (3)$q = 6^2, n = 100, T = 400$. As for the signal level, we vary the spectral norm of $\Lambda$ by $\kappa = 0.5, 0.8, 1$. For each simulation setting, i.e. different combination of Scenario A, C and $(q, n, T, \kappa)$, we conduct 20 experiments.

We summarize the numerical performance of ANH, FAR-F2 and FAR-Tr in Table 3 and in Table 4, where we report the mean MISE (MISE$_{avg}$) and mean PE (PE$_{avg}$) across the 20 experiments. For each experiment, we also calculate the percentage improvement of prediction by RKHS based methods over ANH via Ratio= (PE(ANH) / PE(FAR-x)−1) × 100%. A positive ratio indicates improvement by FAR-F2 or FAR-Tr. We also report the mean ratio (denoted by R$_{avg}$).

| | | Scenario A: $q = 6, n = 36, T = 100$ | | | Scenario A: $q = 6, n = 100, T = 400$ | | |
|---|---|---|---|---|---|---|---|
| | Method | MISE$_{avg}$ | PE$_{avg}$ | R$_{avg}$(%) | MISE$_{avg}$ | PE$_{avg}$ | R$_{avg}$(%) |
| $\kappa = 0.5$ | FAR-F2 | **0.835** | **2.445**$\times10^{-2}$ | 11.74 | **0.812** | **2.299** $\times10^{-2}$ | 33.23 |
| | FAR-Tr | 0.835 | $2.455\times10^{-2}$ | 11.28 | 0.812 | $2.298\times10^{-2}$ | 33.29 |
| | ANH | 1.573 | $2.732\times10^{-2}$ | | 1.003 | $3.063\times10^{-2}$ | |
| $\kappa = 0.8$ | FAR-F2 | **0.833** | **2.403**$\times10^{-2}$ | 74.78 | **0.803** | **2.381**$\times10^{-2}$ | 179.88 |
| | FAR-Tr | 0.833 | $2.405\times10^{-2}$ | 74.64 | 0.803 | $2.381\times10^{-2}$ | 179.88 |
| | ANH | 1.216 | $4.200\times10^{-2}$ | | 0.923 | $6.664\times10^{-2}$ | |
| | | Scenario A: $q = 6^2, n = 36, T = 400$ | | | Scenario A: $q = 6^2, n = 100, T = 400$ | | |
| | Method | MISE$_{avg}$ | PE$_{avg}$ | R$_{avg}$(%) | MISE$_{avg}$ | PE$_{avg}$ | R$_{avg}$(%) |
| $\kappa = 0.5$ | FAR-F2 | **0.964** | **0.159** | 19.50 | **0.963** | **0.147** | 28.57 |
| | FAR-Tr | 0.964 | 0.162 | 17.28 | 0.964 | 0.147 | 28.57 |
| | ANH | 0.967 | 0.190 | | 0.999 | 0.189 | |
| $\kappa = 0.8$ | FAR-F2 | 0.966 | 0.158 | 71.52 | **0.962** | **0.161** | 188.20 |
| | FAR-Tr | 0.965 | **0.157** | 72.61 | 0.963 | 0.161 | 188.20 |
| | ANH | **0.961** | 0.271 | | 0.993 | 0.464 | |

Table 3: Bold font indicates the best performance.

|  |  | Scenario C: $q = 6, n = 36, T = 100$ |  | Scenario C: $q = 6, n = 100, T = 400$ |  |
|---|---|---|---|---|---|
|  | Method | MISE$_{avg}$ | PE$_{avg}$ | MISE$_{avg}$ | PE$_{avg}$ |
| $\kappa = 0.5$ | FAR-F2 | 0.892 | $2.381\times10^{-2}$ | 0.576 | $2.361\times10^{-2}$ |
|  | FAR-Tr | **0.597** | $\mathbf{2.341}\times10^{-2}$ | **0.449** | $\mathbf{2.341}\times10^{-2}$ |
|  | ANH | 3.705 | $2.350\times10^{-2}$ | 2.824 | $2.342\times10^{-2}$ |
| $\kappa = 0.8$ | FAR-F2 | 0.436 | $2.297\times10^{-2}$ | 0.247 | $2.394\times10^{-2}$ |
|  | FAR-Tr | **0.329** | $\mathbf{2.252}\times10^{-2}$ | **0.208** | $\mathbf{2.382}\times10^{-2}$ |
|  | ANH | 1.414 | $2.268\times10^{-2}$ | 2.512 | $2.382\times10^{-2}$ |
|  |  | Scenario C: $q = 6^2, n = 36, T = 400$ |  | Scenario C: $q = 6^2, n = 100, T = 400$ |  |
|  | Method | MISE$_{avg}$ | PE$_{avg}$ | MISE$_{avg}$ | PE$_{avg}$ |
| $\kappa = 0.8$ | FAR-F2 | 0.715 | 0.157 | 0.822 | 0.140 |
|  | FAR-Tr | **0.711** | **0.156** | **0.819** | **0.140** |
|  | ANH | 1.002 | 0.156 | 3.163 | 0.140 |
| $\kappa = 1$ | FAR-F2 | 0.777 | **0.160** | 0.839 | 0.148 |
|  | FAR-Tr | **0.775** | **0.160** | **0.837** | 0.148 |
|  | ANH | 1.002 | **0.160** | 2.062 | **0.146** |

Table 4: Numerical performance in scenario C for FAR(1) processes. Bold font indicates the best performance.

In scenario A, the two RKHS based method do not have much difference in terms of error on operator A, and the Frobenius penalty estimator is slightly better than nuclear penalty estimator in terms of prediction error. They outperform the fPCA method in both criteria.

In the low rank setting, although the improvement on the prediction error is not quite clear, the nuclear penalty estimator has the lowest estimation error among these 3 methods.

# 6 Conclusions and Future Work

This paper studies the estimation and prediction of a functional autoregressive process and extends the inference framework from one dimensional to multidimensional via the tools of Reproducing Kernel Hilbert Spaces (RKHS) of functional autoregressive time series. We work on the complexity and regularization to develop a computationally efficient algorithm. The paper also extend functional PCA method to multidimensional scenario, which serve as a baseline method for simulations. Future work could be applying this method to ARGO float data, which collects the temperature, salinity and other measurements pertaining to the biology/chemistry of the ocean.

# References

Aue, A., Norinho, D. D., and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110(509):378–392.

Besse, P. C. and Cardot, H. (1996). Approximation spline de la prevision d'un processus fonctionnel autorégressif d'ordre 1. *Canadian Journal of Statistics*, 24(4):467–487.

Besse, P. C., Cardot, H., and Stephenson, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27(4):673–687.

Bosq, D. (2000). *Linear processes in function spaces: theory and applications.* Springer-Verlag New York.

Didericksen, D., Kokoszka, P., and Zhang, X. (2012). Empirical properties of forecasts with the functionalautoregressive model. *Computational Statistics*, (27):285–298.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis.* Springer-Verlag New York.

Gu, C. (2013). *Smoothing Spline ANOVA Models.* Springer-Verlag New York, 2 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning.* Springer-Verlag New York, 2 edition.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer.

Hyndman, R. J. and Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3):199–211.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *Annals of Statistics*, 38(3):1845–1884.

Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, page 457–464.

Lee, D.-J. and Durbán, M. (2011). P-spline anova-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, 11(1):49–69.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis.* Springer-Verlag New York.

Shang, H. L. (2013). Functional time series approach for forecasting very short-term electricity demand. *Journal of Applied Statistics*, 40(1):152–168.

Sun, X., Du, P., Wang, X., and Ma, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Wang, D., Zhao, Z., Willett, R., and Yau, C. Y. (2020a). Functional autoregressive processes in reproducing kernel hilbert spaces. *arXiv preprint arXiv:2011.13993*.

Wang, D., Zhao, Z., Yu, Y., and Willet, R. (2020b). Functional linear regression with mixed predictors. *arXiv preprint arXiv:2012.00460*.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Annals of Statistics*, 38(6):3412–3444.