Statistica Sinica Preprint No: SS-2024-0282	
Title	Network Functional Linear Regression
Manuscript ID	SS-2024-0282
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0282
Complete List of Authors	Xingyu Yan and Yanyuan Ma
Corresponding Authors	Xingyu Yan
E-mails	yan@jsnu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Network Functional Linear Regression

Xingyu Yan and Yanyuan Ma

Jiangsu Normal University and Pennsylvania State University

Abstract: We consider a large-scale network where a scalar response and a functional predictor are observed for each individual. To incorporate the network information and to depict the dynamic impact of the functional predictor on the response of each individual, we investigate a network functional linear model. The model assumes that each individual's response can be explained by a linear combination of the responses of the neighbors and a functional regression of the individual. We first approximate functional regression coefficient by a finite representation based on functional principal component analysis technique (FPCA) and then introduce a novel least-squares type of procedure to estimate the coefficients after dimension reduction. In addition, we introduce two modified BIC-type criteria for choosing the number of principal components. We study the convergence rates of the functional regression coefficients and establish the asymptotic normality of the network autoregression coefficients, as well as the consistency of the model selection procedures. Extensive simulation studies are conducted to evaluate the finite sample performance of our proposed method. Finally, we illustrate the usefulness of our method by applying it to two applications.

Key words and phrases: Functional data, Functional linear regression, Functional principal components, Least squares estimator, Network.

1. Introduction

In functional data analysis, functional linear regression (FLR) is often used to depict the association between a functional predictor and a real-valued response. Specifically, for $i = 1, \ldots, n$, let Y_i be the scalar response and $Z_i(t)$ be the associated functional predic-

tor. Here, $Z_i(t)$'s are considered as independent realizations of an underlying stochastic process $\{Z(t), t \in \mathcal{I}\}$, where $\mathcal{I}$ denotes the interval on which the stochastic process is defined. The classical functional linear regression model takes the form

$$Y_i = \alpha + \int_{\mathcal{I}} Z_i(t)\beta(t)dt + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1.1}$$

Here the intercept α and the errors ε_i are scalars and the slope $\beta(\cdot)$ is a function. Model (1.1) has been successfully applied to a wide range of applications especially in the clinical, biometrical, epidemiological, social and economic studies. For a comprehensive summary on the FLR and its applications, we refer to Ferraty and Vieu (2006); Kokoszka and Reimherr (2017).

There have been various works on extending (1.1). For example, Yao et al. (2005) considered the functional response and investigated the corresponding asymptotic properties. Spline functions representation and minimax convergence rate were considered by Crambes et al. (2009) and Cai and Yuan (2012), respectively. Wang et al. (2017) developed a class of generalized scalar-on-image regression models via total variation. Chen et al. (2022) further added dependent error on the functional predictor. Despite the remarkable progress, almost all forms of FLR assume that the individuals are independently sampled.

On the other hand, network structure information is becoming increasingly available due to the improved capability to record and store a wide variety of data and advances in scientific computing. Accordingly, it becomes feasible to obtain the relation between individuals (nodes) in a network. Therefore, rich information is available regarding both the individual attributes and network structure, and the challenge lies in finding appropriate statistical methods to incorporate and utilize both sources of information. Moreover, in many modern

applications including our motivating data examples, individuals not only exhibit network dependence structure but also possess functional predictor. For instance, the crop yield at the county level in USA, with spatial correlation, inherently generates a large amount of network data. Since measurements of meteorological variables, particularly temperature and precipitation, are available on a daily basis and their effects on yield vary at different growing stages of the crop, it is natural to treat them as functional predictors (Park et al., 2023). As illustrated earlier, current FLR literature largely relies on the assumption that individuals are independently sampled. Thus, existing models and methods are no longer suitable for addressing this type of datasets. Therefore, how to incorporate network structure within the framework of functional regression and conduct model estimation effectively remains an open problem for research. In this article, we attempt to fill this gap by developing estimation procedure for a network version of FLR and providing corresponding theoretical guarantees.

In order to incorporate the network dependence structure into functional regression modeling, we employ the concept of the network autoregression model. In particular, we consider a network with n nodes, indexed with $i = 1, \ldots, n$. To describe the network structure, we define an adjacency matrix $\boldsymbol{\mathcal{A}} = (a_{ij}) \in \mathbb{R}^{n \times n}$, and let $a_{ij} = 1$ if there is a link from the ith node to the jth node; otherwise, $a_{ij} = 0$. For the ith node, the response is recorded as Y_i, and the associated functional predictor, denoted by $Z_i(t)$, is considered as an independent realization of an underlying stochastic process. Within the network framework, Y_i might be influenced by three different factors. The first factor is node-specific functional predictor $Z_i(t)$, which is consistent with the classical functional linear regression models. The second factor is the overall response from the node's connected individuals, represented as $\sum_{j=1}^{n} a_{ij} Y_j$, which is referred to as the network autoregression term and quantifies the

network influence on the node's response. Lastly, the unexplained variation is attributed to independent random noise.

To our best knowledge, the current network regression literature mainly adopts classic network autoregression model and focuses on vector-valued covariates. (e.g., Huang et al., 2019; Lee et al., 2010; Zhu et al., 2020; Le and Li, 2022; Tho et al., 2023; Ren et al., 2024). Hence they cannot be directly used in modeling the network with functional data. In addition, the theoretical analysis is challenging since the functional data are usually defined on a space that is intrinsically infinite-dimensional. There are a few pioneering works exploring network autoregression under different functional regression settings. Of particular relevance to the present work are the works of Zhu et al. (2017, 2022). Specifically, Zhu et al. (2017) developed a social network vector autoregression model, which takes network structure into account when modeling the dynamic pattern of an ultra-high dimensional vector response. They further illustrated the usefulness of the model through a social network dataset. Zhu et al. (2022) proposed a network functional varying coefficient model and devised a nonparametric least-squares type estimator, which is feasible when the responses are functional data. Although these studies share a similarity with the present work, they do not address the case of functional predictor.

Compared to the existing literature, our innovation and contribution are in three perspectives. First, we provide a novel network functional linear model that characterizes the dynamic association between functional predictor and scalar-type response, which incorporates valuable network information from data. The developed model adds to the richness of both the classical network autoregression models and the functional regression models. Second, to reduce the computational burden, especially for large-scale network, combining

FPCA technique, we propose a novel utilization of the composite least squares method to estimate the proposed model. The estimator takes advantage of the sparsity of the network structure to reduce the computational burden. Third, rigorous asymptotic theory is established. We allow the number of principal components in the model to diverge to infinity with the sample size, to acknowledge the fact that functional data reside in infinite-dimensional space. Our theory is fundamentally different from those in the vector network autoregression model literature, since our predictor in the model are estimated FPCA scores that are contaminated with measurement errors. To the best of our knowledge, this is the first theoretical study of the functional network regression under the scalar-on-function setting.

The rest of the article is organized as follows. In Section 2, we first present the model for functional linear regression with available network information and develop the novel least squares estimator for analyzing networks based on the proposed model. We then introduce some necessary notations. At the end of Section 2, we develop two criteria to select the number of principal components. In Section 3, we investigate the asymptotic properties of our proposed estimators for the functional regression coefficient and the network autocorrelation coefficient, as well as the consistency of the model selection procedures. Section 4 presents the finite sample performance of the method through a series of simulation studies. In Section 5, two real data examples are provided for illustration. They are the studies of county-level corn yield in five Midwestern states in the United States and air pollution pattern (especially PM2.5) in United Kingdom. In Section 6, we summarize our article and comment on some interesting future research directions. Auxiliary lemmas, additional simulation results, theoretical analysis, and all proofs are provided in the Supplementary Material for clarity.

2. Model and Estimation

2.1 Model and eigenbasis representation

We begin by reviewing the fundamental technique used for the dimensionality reduction and regularization of functional data, known as FPCA. Recall that $\{Z(t), t \in \mathcal{I}\}$ is an L^2 stochastic process on an interval $\mathcal{I}$. The mean and covariance functions of $Z(t)$ are, respectively, $E\{Z(t)\} = \mu(t)$ and $G(s,t) = \text{cov}(Z(s), Z(t))$. FPCA implies that there exists a sequence of orthonormal functions $\{\phi_k(t)\}_{k=1,2,...,}$ which form a complete orthonormal basis of the functional space, with associated nonnegative and nonincreasing eigenvalues $\{\lambda_k\}_{k=1,2,...,}$. Predictor processes can then be represented by the Karhunen-Loève (K-L) expansion $Z(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$, where $\xi_k = \int_{\mathcal{I}} \{Z(t) - \mu(t)\}\phi_k(t)dt$. The random variables ξ_k are the functional principal components, also referred to as FPC scores. These scores are uncorrelated and satisfy $E(\xi_k) = 0$ and $\text{var}(\xi_k) = \lambda_k$.

For $i = 1, \ldots, n$, let functional predictor $Z_i(t)$ be identically and independently distributed (i.i.d.) copies of $Z(t)$ with a scalar response Y_i. Without loss of generality, we can assume that $\mu(t) = 0$ and $EY_i = 0$. Then the network functional linear model proposed is introduced as follows

$$Y_i = \int_{\mathcal{I}} Z_i(t)\beta(t)dt + \rho \sum_{j=1}^{n} w_{ij} Y_j + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2}$$

The first term $\int_{\mathcal{I}} Z_i(t)\beta(t)dt$ characterizes the nodal effect from its own functional feature $Z_i(t)$ as in classical FLR (1.1). Here $\beta(\cdot)$ represents the functional regression coefficient that captures time varying effects. The second term $\rho \sum_{j=1}^{n} w_{ij} Y_j$ is the network effect, which describes the influence from the connected neighbors of node i, where ρ can be intuitively

interpreted as a measure of the strength of the network dependence (i.e., the so-called network autocorrelation coefficient). We assume $|\rho| < 1$ throughout the article. The matrix $(w_{ij})_{i=1,\ldots,n;j=1,\ldots,n}$ is row-normalized adjacency matrix with a zero diagonal, describing the network structure defined in Zhu et al. (2017), and we assume that its off-diagonal elements are nonnegative. The last term ε_i denotes random error.

We expand $\beta(t)$ in terms of the eigenfunctions of predictor process $Z(t)$ such that $\beta(t) = \sum_{k=1}^{\infty} b_k \phi_k(t)$, and by K-L expansion discussed above, the original model (2.2) can be expressed equivalently as

$$Y_i = \sum_{k=1}^{\infty} \xi_{ik} b_k + \rho \sum_{j=1}^{n} w_{ij} Y_j + \varepsilon_i. \tag{2.3}$$

It is commonly assumed that the conditional distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ given $Z(\cdot)$ only depends on $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^{\mathrm{T}}$ (e.g., Li et al., 2010; Kong et al., 2016),

$$Y_i = \sum_{k=1}^{K} \xi_{ik} b_k + \rho \sum_{j=1}^{n} w_{ij} Y_j + \varepsilon_i. \tag{2.4}$$

The number of included components K, typically chosen simply as the smallest number of components that explain a large enough fraction of the overall variance of predictor processes. We may treat K as a regularization parameter that balances approximation bias and estimation variance, and we assume K diverges with sample size n. For ease of presentation, we define the following notations. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \in \mathbb{R}^n$ be the response vector, $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n \times n}$, and $\mathbf{A} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times K}$, $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^{\mathrm{T}}$. In addition, define the coefficient vector as $\mathbf{b} = (b_1, \ldots, b_K)^{\mathrm{T}} \in \mathbb{R}^K$ and noise vector as $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}} \in \mathbb{R}^n$. Specifically, $\boldsymbol{\varepsilon}$ is assumed to satisfy $E(\boldsymbol{\varepsilon}|\mathbf{A}) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$, where $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ is an

identity matrix. Note that we do not require normality on $\boldsymbol{\varepsilon}$. Then, model (2.4) can be expressed in matrix form as follows

$$\mathbf{Y} = \mathbf{Ab} + \rho\mathbf{WY} + \boldsymbol{\varepsilon}. \tag{2.5}$$

Further, let $\mathbf{S}(\rho) = \mathbf{I} - \rho\mathbf{W}$, then model (2.5) leads to

$$\mathbf{Y} = \mathbf{S}^{-1}(\rho)\mathbf{Ab} + \mathbf{S}^{-1}\boldsymbol{\varepsilon}. \tag{2.6}$$

Obviously, the mean and covariance of response $\mathbf{Y}$ are $\mathbf{S}^{-1}(\rho)\mathbf{Ab}$ and $\sigma^2\mathbf{S}^{-1}(\rho)(\mathbf{S}^{-1}(\rho))^{\mathrm{T}}$, respectively. Our estimation procedure is to borrow the idea of composite likelihood based on model (2.6). We describe the detailed approach in the next section.

2.2 Estimation procedure

FPC scores usually cannot be observed directly. Therefore we need firstly to estimate FPC scores from the observed functional data. Specifically, obtain estimator $\widehat{G}(s,t)$ for the $G(s,t)$ by moment estimation, that is $\widehat{G}(s,t) = n^{-1}\sum_{i=1}^{n} Z_i(s)Z_i(t)$. In this article, we focus on the case where predictor trajectories are recorded on equally spaced grid points. In practice, if the trajectories are not fully observed for each subject, and instead, they are recorded at a series of different time points, and then one can employ local polynomial smoothing to smooth the trajectories before proceeding to the next step. We refer to Kong et al. (2016) for more detailed discussions. Implement FPCA to estimate eigenvalues λ_k, eigenfunction $\phi_k(t)$ and the FPC scores ξ_{ik}, $k = 1, \ldots K$. Denote the corresponding estimators $\widehat{\lambda}_k$, $\widehat{\phi}_k(t)$ and $\widehat{\xi}_{ik}$ respectively. They satisfy $\int Z_i(t)\widehat{\phi}_k(t)dt = \widehat{\xi}_{ik}$, $\int \widehat{\phi}_k^2(t)dt = 1$, $n^{-1}\sum_{i=1}^{n}\widehat{\xi}_{ik} = 0$,

$n^{-1}\sum_{i=1}^{n}\widehat{\xi}_{ik}^2 = \widehat{\lambda}_k$ and $n^{-1}\sum_{i=1}^{n}\widehat{\xi}_{ik}\widehat{\xi}_{i\ell} = 0$ for $\ell \neq k$. See Supplementary Material for detailed verification. Subsequently, the regression coefficients $\{b_k\}_{k=1,\ldots,K}$ and ρ can be analyzed below. The sample version of model (2.4) can be further approximated by

$$Y_i \approx \sum_{k=1}^{K}\widehat{\xi}_{ik}b_k + \rho\sum_{j=1}^{n}w_{ij}Y_j + \varepsilon_i. \tag{2.7}$$

Likewise, in matrix notation, the model (2.7) can be written as

$$\mathbf{Y} \approx \widehat{\mathbf{A}}\mathbf{b} + \rho\mathbf{W}\mathbf{Y} + \boldsymbol{\varepsilon}, \tag{2.8}$$

where the estimated scores $\widehat{\mathbf{A}} = (\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times K}$ serve as the predictor variables, where $\widehat{\boldsymbol{\xi}}_i = (\widehat{\xi}_{i1}, \ldots, \widehat{\xi}_{iK})^{\mathrm{T}}$. Note that $n^{-1}\widehat{\mathbf{A}}^{\mathrm{T}}\widehat{\mathbf{A}} = \widehat{\boldsymbol{\Lambda}} \equiv \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K)$. Thus, one can obtain the least squares estimator of $(\mathbf{b}, \rho)$ by minimizing

$$\frac{1}{n}\left\|\mathbf{Y} - \mathbf{S}^{-1}(\rho)\widehat{\mathbf{A}}\mathbf{b}\right\|^2,$$

where $\|\cdot\|$ denotes the L_2 norm. However, the above estimation equation can be computationally expensive because of taking inverse of the large matrix $\mathbf{S}(\rho)$. To avoid taking inverse of $\mathbf{S}(\rho)$, we consider the objective function

$$Q(\mathbf{b}, \rho) = \frac{1}{n}\left\|\mathbf{S}(\rho)\mathbf{Y} - \widehat{\mathbf{A}}\mathbf{b}\right\|^2. \tag{2.9}$$

But after some calculations, we observe that $E\{\partial Q(\mathbf{b}, \rho)/\partial\rho\} \neq 0$ because the expression of $\partial Q(\mathbf{b}, \rho)/\partial\rho$ involves a quadratic form with respect to $\mathbf{Y}$, whose expectation is nonzero.

More detailed explanation can be found in Supplementary Material S3.1. Therefore, we introduce the following procedure. We consider a working model of (2.8) where $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, resulting in $\mathbf{Y}|\widehat{\mathbf{A}} \sim \mathcal{N}_n\{\mathbf{S}^{-1}(\rho)\widehat{\mathbf{A}}\mathbf{b}, \sigma^2\mathbf{S}^{-1}(\rho)(\mathbf{S}^{-1}(\rho))^{\mathrm{T}}\}$. Let $\mathbf{M}(\rho) = \mathbf{S}(\rho)^{\mathrm{T}}\mathbf{S}(\rho)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}} = \mathbf{S}^{-1}(\rho)\widehat{\mathbf{A}}\mathbf{b}$, then

$$E\left(Y_i|\mathbf{Y}_{-i}, \widehat{\mathbf{A}}\right) = \mu_i - \mathrm{M}_{ii}^{-1}(\rho)\mathbf{M}_{i,-i}(\rho)(\mathbf{Y}_{-i} - \boldsymbol{\mu}_{-i}),$$

where $\mathbf{Y}_{-i}$ is $\mathbf{Y}$ without the ith element, $\mathbf{M}_{i,-i}(\rho)$ is the ith row of $\mathbf{M}(\rho)$ without the ith entry, $\mathrm{M}_{ii}(\rho)$ is ith diagonal entry of matrix $\mathbf{M}(\rho)$, $\boldsymbol{\mu}_{-i}$ is $\boldsymbol{\mu}$ without the ith entry.

For a given $\mathbf{b}$, the composite least squares objective function can be written as

$$Q_1(\widehat{\mathbf{A}}; \mathbf{b}, \rho) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - E\left(Y_i|\mathbf{Y}_{-i}, \widehat{\mathbf{A}}\right)\right\}^2. \tag{2.10}$$

Furthermore,

$$\begin{aligned} Y_i - E\left(Y_i|\mathbf{Y}_{-i}, \widehat{\mathbf{A}}\right) &= Y_i - \mu_i + \mathrm{M}_{ii}^{-1}(\rho)\mathbf{M}_{i,-i}(\rho)\left(\mathbf{Y}_{-i} - \boldsymbol{\mu}_{-i}\right) \\ &= \mathrm{M}_{ii}^{-1}(\rho)\mathbf{M}_{i,\cdot}(\rho)(\mathbf{Y} - \boldsymbol{\mu}) \\ &= \mathbf{e}_i^{\mathrm{T}}\{\mathrm{diag}\mathbf{M}(\rho)\}^{-1}\mathbf{S}(\rho)^{\mathrm{T}}\left(\mathbf{S}(\rho)\mathbf{Y} - \widehat{\mathbf{A}}\mathbf{b}\right), \end{aligned}$$

where $\mathbf{e}_i$ denotes the $n \times 1$ vector of zeros except the ith element which is one. Hence, we have

$$Q_1(\widehat{\mathbf{A}}; \mathbf{b}, \rho) = \frac{1}{n}\left\|\{\mathrm{diag}\mathbf{M}(\rho)\}^{-1}\mathbf{S}(\rho)^{\mathrm{T}}\left(\mathbf{S}(\rho)\mathbf{Y} - \widehat{\mathbf{A}}\mathbf{b}\right)\right\|^2.$$

Remark 1. *Note that* $E\{\partial Q_1(\widehat{\mathbf{A}}; \mathbf{b}, \rho)/\partial\rho\} = 0$ *holds true regardless of the validity of the working model. Therefore, the assumption of normality on the noise is solely used to facilitate the construction of the objective function. The validity of the methodology and the theoretical results we will develop later do not rely on the assumption of normality. Similar descriptions can be found in Huang et al. (2019) and Zhu et al. (2020).*

In summary, we state the three-step algorithm as follows.

Step 1: Implement FPCA to estimate the FPC scores $\{\xi_{i1}, \ldots, \xi_{iK}\}$ of Z_i, where K is selected as described in Section 2.3.

Step 2: At any ρ, obtain the estimator $\widehat{\mathbf{b}}(\rho)$ through solving

$$\begin{aligned} \widehat{\mathbf{b}}(\rho) &= \arg\min_{\mathbf{b}} Q(\mathbf{b}, \rho) \\ &= \arg\min_{\mathbf{b}} \frac{1}{n} \left\| \mathbf{S}(\rho)\mathbf{Y} - \widehat{\mathbf{A}}\mathbf{b} \right\|^2 \\ &= \widehat{\mathbf{\Lambda}}^{-1} \left(\frac{1}{n} \widehat{\mathbf{A}}^{\mathrm{T}} \mathbf{S}(\rho)\mathbf{Y} \right). \end{aligned} \tag{2.11}$$

Step 3: Obtain the estimator $\widehat{\rho}$ by

$$\begin{aligned} \widehat{\rho} &= \arg\min_{\rho} Q_1 \left\{ \widehat{\mathbf{A}}; \widehat{\mathbf{b}}(\rho), \rho \right\} \\ &= \arg\min_{\rho} \frac{1}{n} \left\| \{\mathrm{diag}\mathbf{M}(\rho)\}^{-1} \mathbf{S}(\rho)^{\mathrm{T}} \left\{ \mathbf{S}(\rho)\mathbf{Y} - \widehat{\mathbf{A}}\widehat{\mathbf{b}}(\rho) \right\} \right\|^2. \end{aligned} \tag{2.12}$$

Before concluding this subsection, we introduce some notations which will be used repetitively in subsequent derivations. Let $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ and $\mathbf{A} = (\langle Z_i, \phi_k \rangle) \in \mathbb{R}^{n \times K}$. Let

$$b_k(\rho) = \lambda_k^{-1} \int \phi_k(t) E \left\{ \left(Y_i - \rho \sum_{j=1}^{n} w_{ij} Y_j \right) Z_i(t) \right\} dt,$$

and $\mathbf{b}(\rho) = \{b_1(\rho), \ldots, b_K(\rho)\}^{\mathrm{T}}$. It follows then that

$$\frac{\partial b_k(\rho)}{\partial \rho} = -\lambda_k^{-1} \int \phi_k(t) E \left\{ \sum_{j=1}^{n} w_{ij} Y_j Z_i(t) \right\} dt, \tag{2.13}$$

which does not depend on ρ. Let $\mathbf{D}(\rho) = \{\mathrm{diag}\mathbf{M}(\rho)\}^{-1}$, and by (2.12), taking derivative of $Q_1\{\widehat{\mathbf{A}}; \widehat{\mathbf{b}}(\rho), \rho\}$ with respect to ρ, we define, for any ρ^*,

$$\begin{aligned} \frac{\mathrm{d}Q_1\{\widehat{\mathbf{A}}; \widehat{\mathbf{b}}(\rho), \rho\}}{\mathrm{d}\rho}\Big|_{\rho=\rho^*} &= \left[\frac{\partial Q_1\{\widehat{\mathbf{A}}; \widehat{\mathbf{b}}(\rho), \rho\}}{\partial \rho} + \frac{\partial Q_1\{\widehat{\mathbf{A}}; \widehat{\mathbf{b}}(\rho), \rho\}}{\partial \widehat{\mathbf{b}}(\rho)^{\mathrm{T}}} \frac{\partial \widehat{\mathbf{b}}(\rho)}{\partial \rho} \right] \Big|_{\rho=\rho^*} \\ &= \frac{1}{n} \sum_{i=1}^{n} S_i\{\widehat{\mathbf{A}}; \widehat{\mathbf{b}}(\rho^*), \rho^*\}, \end{aligned} \tag{2.14}$$

where $n^{-1} \sum_{i=1}^{n} S_i\{\widehat{\mathbf{A}}; \widehat{\mathbf{b}}(\rho^*), \rho^*\}$ is obtained from replacing $\mathbf{b}$, ρ and $1/\sqrt{n}$ in following formula (2.15) with $\widehat{\mathbf{b}}$, ρ^* and $1/n$, respectively,

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\} \\ = \; & \frac{1}{\sqrt{n}} \left\{ \mathbf{D}(\rho)\mathbf{M}(\rho)\mathbf{Y} - \mathbf{D}(\rho)\mathbf{S}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}} \mathbf{b}(\rho) \right\}^{\mathrm{T}} \\ & \times \left[\left\{ \frac{\partial \mathbf{D}(\rho)}{\partial \rho} \mathbf{M}(\rho) + \mathbf{D}(\rho) \frac{\partial \mathbf{M}(\rho)}{\partial \rho} \right\} \mathbf{Y} - \left\{ \frac{\partial \mathbf{D}(\rho)}{\partial \rho} \mathbf{S}(\rho)^{\mathrm{T}} + \mathbf{D}(\rho) \left(-\mathbf{W}^{\mathrm{T}}\right) \right\} \widehat{\mathbf{A}} \mathbf{b}(\rho) \right. \\ & \left. - \mathbf{D}(\rho)\mathbf{S}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}} \frac{\partial \mathbf{b}(\rho)}{\partial \rho} \right], \end{aligned} \tag{2.15}$$

where

$$\begin{aligned} S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\} \equiv & \left[M_{ii}^{-1}(\rho) \left\{ \sum_{j=1}^{n} M_{ij}(\rho) Y_j - \sum_{k=1}^{K} \left(\widehat{\xi}_{ik} - \rho \sum_{j=1}^{n} w_{ji} \widehat{\xi}_{jk} \right) b_k(\rho) \right\} \right] \\ & \times \left(\frac{dM_{ii}^{-1}(\rho)}{d\rho} \left\{ \sum_{j=1}^{n} M_{ij}(\rho) Y_j - \sum_{k=1}^{K} \left(\widehat{\xi}_{ik} - \rho \sum_{j=1}^{n} w_{ji} \widehat{\xi}_{jk} \right) b_k(\rho) \right\} \right. \end{aligned}$$

$$+ M_{ii}^{-1}(\rho)\left[\sum_{j=1}^{n}\frac{dM_{ij}(\rho)}{d\rho}Y_j - \sum_{k=1}^{K}\left\{\left(-\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right)b_k(\rho) + \left(\widehat{\xi}_{ik} - \rho\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right)\frac{db_k(\rho)}{d\rho}\right\}\right], \tag{2.16}$$

where $M_{ii}(\rho) = 1 + \rho^2\sum_{j=1}^{n}w_{ji}^2 - 2\rho w_{ii}$, $M_{ij}(\rho) = -\rho w_{ji} - \rho w_{ij} + \rho^2\sum_{s=1}^{n}w_{si}w_{sj}$ for $i \neq j$.

Define the partial derivative of $S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\}$ with respect to $\mathbf{b}(\rho)$ as follows,

$$\frac{\partial S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\}}{\partial \mathbf{b}(\rho)^{\mathrm{T}}} = \left[\begin{array}{ccc} \frac{\partial S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\}}{\partial b_1(\rho)} & \cdots & \frac{\partial S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\}}{\partial b_K(\rho)} \end{array}\right],$$

where

$$\begin{aligned}
&\frac{\partial S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\}}{\partial b_k(\rho)} \\
=\ & -2M_{ii}^{-1}(\rho)\left(\widehat{\xi}_{ik} - \rho\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right) \\
& \times\left(\frac{dM_{ii}^{-1}(\rho)}{d\rho}\left\{\sum_{j=1}^{n}M_{ij}(\rho)Y_j - \sum_{k=1}^{K}\left(\widehat{\xi}_{ik} - \rho\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right)b_k(\rho)\right\}\right) \\
& + \left\{-M_{ii}^{-2}(\rho)\left(\widehat{\xi}_{ik} - \rho\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right)\right\} \\
& \times\left[\sum_{j=1}^{n}\frac{dM_{ij}(\rho)}{d\rho}Y_j - \sum_{k=1}^{K}\left\{\left(-\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right)b_k(\rho) + \left(\widehat{\xi}_{ik} - \rho\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right)\frac{db_k(\rho)}{d\rho}\right\}\right] \\
& + \left\{M_{ii}^{-2}(\rho)\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right\}\left\{\sum_{j=1}^{n}M_{ij}(\rho)Y_j - \sum_{k=1}^{K}\left(\widehat{\xi}_{ik} - \rho\sum_{j=1}^{n}w_{ji}\widehat{\xi}_{jk}\right)b_k(\rho)\right\}.
\end{aligned} \tag{2.17}$$

According to (2.17),

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{\partial S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\}}{\partial \mathbf{b}(\rho)^{\mathrm{T}}} \\
=\ & \frac{1}{n}\left[-2\left\{\mathbf{Y}^{\mathrm{T}}\mathbf{M}(\rho) - \mathbf{b}(\rho)^{\mathrm{T}}\widehat{\mathbf{A}}^{\mathrm{T}}\mathbf{S}(\rho)\right\}\left\{\frac{\partial \mathbf{D}(\rho)}{\partial\rho}\mathbf{D}(\rho)\mathbf{S}(\rho)^{\mathrm{T}}\widehat{\mathbf{A}}\right\}\right.
\end{aligned}$$

$$
\begin{aligned}
&- \left\{ \mathbf{Y}^{\mathrm{T}} \frac{\partial \mathbf{M}(\rho)}{\partial \rho} + \mathbf{b}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}}^{\mathrm{T}} \mathbf{W} - \frac{\partial \mathbf{b}(\rho)^{\mathrm{T}}}{\partial \rho} \widehat{\mathbf{A}}^{\mathrm{T}} \mathbf{S}(\rho) \right\} \left\{ \mathbf{D}(\rho)^2 \mathbf{S}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}} \right\} \\
&+ \left\{ \mathbf{Y}^{\mathrm{T}} \mathbf{M}(\rho) - \mathbf{b}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}}^{\mathrm{T}} \mathbf{S}(\rho) \right\} \left\{ \mathbf{D}(\rho)^2 \mathbf{W}^{\mathrm{T}} \widehat{\mathbf{A}} \right\} \Big]. \qquad (2.18)
\end{aligned}
$$

Taking derivative of (2.15) with respect to ρ, by $\partial^2 b_k(\rho)/\partial \rho^2 = 0$, we have

$$
\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{d} S_i\{\widehat{\mathbf{A}}; \mathbf{b}(\rho), \rho\}}{\mathrm{d}\rho} \\
= \;& \frac{1}{n} \Bigg[\mathbf{Y}^{\mathrm{T}} \mathbf{M}(\rho) \frac{\partial \mathbf{D}(\rho)}{\partial \rho} + \mathbf{Y}^{\mathrm{T}} \frac{\partial \mathbf{M}(\rho)}{\partial \rho} \mathbf{D}(\rho) - \left\{ \mathbf{b}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}}^{\mathrm{T}} \mathbf{S}(\rho) \frac{\partial \mathbf{D}(\rho)}{\partial \rho} \right. \\
&\quad \left. + \mathbf{b}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}}^{\mathrm{T}} (-\mathbf{W}) \mathbf{D}(\rho) + \frac{\partial \mathbf{b}(\rho)^{\mathrm{T}}}{\partial \rho} \widehat{\mathbf{A}}^{\mathrm{T}} \mathbf{S}(\rho) \mathbf{D}(\rho) \right\} \Bigg] \\
&\times \Bigg[\left\{ \frac{\partial \mathbf{D}(\rho)}{\partial \rho} \mathbf{M}(\rho) + \mathbf{D}(\rho) \frac{\partial \mathbf{M}(\rho)}{\partial \rho} \right\} \mathbf{Y} \\
&\quad - \left\{ \frac{\partial \mathbf{D}(\rho)}{\partial \rho} \mathbf{S}(\rho)^{\mathrm{T}} + \mathbf{D}(\rho) \left(-\mathbf{W}^{\mathrm{T}}\right) \right\} \widehat{\mathbf{A}} \mathbf{b}(\rho) - \mathbf{D}(\rho) \mathbf{S}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}} \frac{\partial \mathbf{b}(\rho)}{\partial \rho} \Bigg] \\
&+ \frac{1}{n} \left\{ \mathbf{Y}^{\mathrm{T}} \mathbf{M}(\rho) \mathbf{D}(\rho) - \mathbf{b}(\rho)^{\mathrm{T}} \widehat{\mathbf{A}}^{\mathrm{T}} \mathbf{S}(\rho) \mathbf{D}(\rho) \right\} \\
&\times \Bigg[\left\{ \frac{\partial^2 \mathbf{D}(\rho)}{\partial \rho^2} \mathbf{M}(\rho) + 2 \frac{\partial \mathbf{D}(\rho)}{\partial \rho} \frac{\partial \mathbf{M}(\rho)}{\partial \rho} + 2 \mathbf{D}(\rho) \mathbf{W}^{\mathrm{T}} \mathbf{W} \right\} \mathbf{Y} \\
&\quad - \left\{ \frac{\partial^2 \mathbf{D}(\rho)}{\partial \rho^2} \mathbf{S}(\rho)^{\mathrm{T}} + 2 \frac{\partial \mathbf{D}(\rho)}{\partial \rho} \left(-\mathbf{W}^{\mathrm{T}}\right) \right\} \widehat{\mathbf{A}} \mathbf{b}(\rho) \\
&\quad - 2 \left\{ \frac{\partial \mathbf{D}(\rho)}{\partial \rho} \mathbf{S}(\rho)^{\mathrm{T}} + \mathbf{D}(\rho) \left(-\mathbf{W}^{\mathrm{T}}\right) \right\} \widehat{\mathbf{A}} \frac{\partial \mathbf{b}(\rho)}{\partial \rho} \Bigg]. \qquad (2.19)
\end{aligned}
$$

2.3 Selection of number of principal components

As we can see from (2.4), the proposed estimation is based on the smoothing parameter K, i.e., the number of principal components. K can be chosen by AIC or BIC type criteria based on the likelihood function, which we will describe in detail below. Let $\widehat{\sigma}_K^2$ be the estimator based on the residuals under K principal components, i.e., $\widehat{\sigma}_K^2 = \mathrm{SSE}_K/n$, where

$\text{SSE}_K = \|\widehat{\mathbf{S}}\mathbf{Y} - \widehat{\mathbf{A}}\widehat{\mathbf{b}}\|^2$, and $\widehat{\mathbf{S}} = \mathbf{I} - \widehat{\rho}\mathbf{W}$. We then define the AIC and BIC criteria as

$$\begin{aligned} \text{AIC}(K) &= \log\left(\widehat{\sigma}_K^2\right) + \frac{2K}{n}, \\ \text{BIC}(K) &= \log\left(\widehat{\sigma}_K^2\right) + \frac{K\log(n)}{n}. \end{aligned}$$

Then the number of principal components is then selected by minimizing the objective functions to obtain $K_{\text{AIC}} = \arg\min_K \text{AIC}(K)$ and $K_{\text{BIC}} = \arg\min_K \text{BIC}(K)$.

The above traditional AIC and BIC criterions are originally devised for fixed predictor dimension. However, the number of principal components K is allowed to diverge. Motivated by the methods of shrinkage tuning parameter selection and the determination of the dimension in context of sufficient dimension reduction (Wang et al., 2009; Zhu et al., 2006), we further develop two procedures of the slightly modified BIC criterion to select K. Specifically, we define

$$\text{BIC}^*(K) = \log\left(\widehat{\sigma}_K^2\right) + K\frac{\log(n)}{n}C_n^*,$$

and

$$\text{BIC}^{**}(K) = \log\left(\widehat{\sigma}_K^2\right) + \frac{KC_n^{**}}{n},$$

where C_n^* and C_n^{**} are some positive constants. In practice, we recommend choosing $C_n^* = \text{loglog}(K^{0.4})$, and C_n^{**} as $n^{0.3}$, respectively, which work very well in our simulations, see Zhu et al. (2006) and Wang et al. (2009) for more related discussion. We will investigate the finite sample performance of these suggested criteria in the simulation studies.

3. Theoretical Properties

For ease of notation, we use C or $C_k, k = 1, 2, \ldots$ to denote generic positive constants. Let $A_n \asymp B_n$ mean $A_n = O_p(B_n)$ and $B_n = O_p(A_n)$. For any $n \times n$ matrix $\mathbf{A}$, let $|\mathbf{A}|$ denote the $\mathbf{A}$ with diagonal elements replaced by its absolute values. Let $(\mathbf{A})_{ij}$ denote the (i, j)th element of $\mathbf{A}$. To investigate the asymptotic properties of the estimators for the functional predictor and the network correlation coefficient, we first present some regularity conditions.

(C1) Define $Y_i^* \equiv Y_i - \rho \sum_{j=1}^n w_{ij} Y_j$ and $g(t) \equiv E\{Y_i^* Z_i(t)\}$. There exists a constant $C > 0$, so that $\int_0^1 E\{Z^4(t)\}dt < C$, $\int_0^1 [E\{Z^4(t)\}]^{1/2} dt < C$, $E(\varepsilon_i^3 | \mathbf{A}) = 0$, $E\varepsilon_i^4 < C$, $E(Y_i^{*4}) < C$, $\int_0^1 \beta^2(t)dt < C$, $E\xi_k^4 < C\lambda_k^2$ for all $k \geqslant 1$. For any $1 \leqslant \ell_a, \ell_b, \ell_c \leqslant n$ and $1 \leqslant k_1, k_2, k_3 \leqslant K$, the third moment $E\left(\xi_{\ell_a k_1} \xi_{\ell_b k_2} \xi_{\ell_c k_3}\right) = 0$. The fourth moment $E(\xi_{jk_1} \xi_{jk_2} \xi_{jk_3}^2) = 0$ for $k_1 \neq k_2 \neq k_3$ and $E(\xi_{jk_1} \xi_{jk_2}^3) = 0$ for $k_1 \neq k_2$.

(C2) For a constant $\alpha > 1$, the eigenvalues λ_k satisfy $C^{-1} k^{-\alpha} \leqslant \lambda_k \leqslant C k^{-\alpha}$, $\quad \lambda_k - \lambda_{k+1} \geqslant C^{-1} k^{-\alpha-1}$ for all $k \geqslant 1$.

(C3) There exists $\tau > \alpha/2 + 1$, $|b_k| \leqslant C k^{-\tau}$ for each $k \geqslant 1$.

(C4) The eigenfunctions $\{\phi_k(t)\}_{k=1}^{\infty}$ are continuous on [0,1] and satisfy $\sup_{k \geqslant 1} \sup_{t \in [0,1]} |\phi_k(t)| = O(1)$.

(C5) There exists a constant $\epsilon > 0$ so that $K = o\{n^{\frac{1}{2(2\alpha+3+\epsilon)}}\}$ and $n^{-1/4} \log K K^{\alpha+2} = O(1)$.

(C6) For all $i = 1, \ldots, n$,

$$\left| \frac{1}{n} \sum_{i=1}^n \frac{\mathrm{d}^2 S_i\{\mathbf{A}; \mathbf{b}(\rho), \rho\}}{\mathrm{d}\rho^2} \right|, \left\| \frac{1}{n} \sum_{i=1}^n \frac{\partial \left[\mathrm{d} S_i\{\mathbf{A}; \mathbf{b}(\rho), \rho\}/\mathrm{d}\rho\right]}{\partial \mathbf{b}(\rho)^{\mathrm{T}}} \right\| \text{ and } \left\| \frac{1}{n} \sum_{i=1}^n \frac{\partial \left[\mathrm{d}^2 S_i\{\mathbf{A}; \mathbf{b}(\rho), \rho\}/\mathrm{d}\rho^2\right]}{\partial \mathbf{b}(\rho)} \right\|$$

are bounded.

(C7) There exists $\tau > \alpha/2 + 1$, $\sup_{\rho\in[-1,1]} |b_k(\rho)| \leqslant Ck^{-\tau}$, $\sup_{\rho\in[-1,1]} |\partial b_k(\rho)/\partial\rho| \leqslant Ck^{-\tau}$ for each $k \geqslant 1$, $\xi_{ik} = O_p(1)$ for any $k = 1, \ldots, K, i = 1, \ldots, n$. Further, $E(n^{-1}\|\mathbf{Y}\|^2) = O(1)$, $\|\mathbf{b}(\rho)\| = O(1)$ and $\|\partial\mathbf{b}(\rho)/\partial\rho\| = O(1)$.

(C8) There exist two positive constants, C_1 and C_2, such that $C_1 \leqslant \lambda_{\min}(n^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{A}) \leqslant \lambda_{\max}(n^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{A}) \leqslant C_2$, $C_1 \leqslant \lambda_{\min}\{\partial^2\mathbf{D}(\rho)/\partial\rho^2\} \leqslant \lambda_{\max}\{\partial^2\mathbf{D}(\rho)/\partial\rho^2\} \leqslant C_2$, $C_1 \leqslant \lambda_{\min}\{\mathbf{M}(\rho)\} \leqslant \lambda_{\max}\{\mathbf{M}(\rho)\} \leqslant C_2$, $C_1 \leqslant \lambda_{\min}\{\partial\mathbf{M}(\rho)/\partial\rho\} \leqslant \lambda_{\max}\{\partial\mathbf{M}(\rho)/\partial\rho\} \leqslant C_2$, $C_1 \leqslant \lambda_{\min}(\mathbf{W}\mathbf{W}^{\mathrm{T}}) \leqslant \lambda_{\max}(\mathbf{W}\mathbf{W}^{\mathrm{T}}) \leqslant C_2$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalue of a matrix.

(C9) The adjacency matrix satisfies $\sup_{i,j} |w_{ij}| = o(1/\sqrt{n})$ for $i, j = 1, \ldots, n$.

(C10)

$$\sqrt{\frac{n}{C_n^* K^{2\alpha+3}\log(n)}}\lim\inf_{n\to\infty}(\min_{k=1,\ldots,K_0}|b_k|) \to \infty, \quad \frac{C_n^* K^{2\alpha+3}\log(n)}{n} \to 0,$$

and $C_n^* \to \infty$, where K_0 denotes the number of the true model.

Remark 2. *Note that assumption* $C_1 \leqslant \lambda_{\min}\{\mathbf{M}(\rho)\} \leqslant \lambda_{\max}\{\mathbf{M}(\rho)\} \leqslant C_2$ *in* (C8) *implies that we can actually find* $C_1 > 0, C_2 > 0$ *so that* $C_1 \leqslant \lambda_{\min}\{\mathbf{M}(\rho)^2\} \leqslant \lambda_{\max}\{\mathbf{M}(\rho)^2\} \leqslant C_2$ *and* $C_1 \leqslant \lambda_{\min}\{\mathbf{S}(\rho)\mathbf{S}(\rho)^{\mathrm{T}}\} \leqslant \lambda_{\max}\{\mathbf{S}(\rho)\mathbf{S}(\rho)^{\mathrm{T}}\} \leqslant C_2$. *Likewise,* $C_1 \leqslant \lambda_{\min}(n^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{A}) \leqslant \lambda_{\max}(n^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{A}) \leqslant C_2$, *and* $C_1 \leqslant \lambda_{\min}(\mathbf{W}\mathbf{W}^{\mathrm{T}}) \leqslant \lambda_{\max}(\mathbf{W}\mathbf{W}^{\mathrm{T}}) \leqslant C_2$ *implie that we can find* $C_1 > 0, C_2 > 0$ *so that* $0 \leqslant \lambda_{\min}(n^{-1}\mathbf{A}\mathbf{A}^{\mathrm{T}}) \leqslant \lambda_{\max}(n^{-1}\mathbf{A}\mathbf{A}^{\mathrm{T}}) \leqslant C_2$, *and* $C_1 \leqslant \lambda_{\min}(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \leqslant \lambda_{\max}(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \leqslant C_2$, *respectively. In addition,* $\lambda_{\max}\{\mathbf{D}(\rho)\}$, $\lambda_{\max}\{|\partial\mathbf{D}(\rho)/\partial\rho|\}$ *and* $\lambda_{\max}\{|\partial^2\mathbf{D}(\rho)/\partial\rho^2|\}$ *are all bounded. This is because the ith diagonal elements of* $\mathbf{D}(\rho)$, $|\partial\mathbf{D}(\rho)/\partial\rho|$ *and* $|\partial^2\mathbf{D}(\rho)/\partial\rho^2|$ *are respectively* $D_{ii}(\rho) = 1/(1+\rho^2\sum_{j=1}^n w_{ji}^2) \leqslant 1$, $|\partial D_{ii}(\rho)/\partial\rho| =$

$2|\rho|\sum_{j=1}^n w_{ji}^2/(1+\rho^2\sum_{j=1}^n w_{ji}^2)^2 \leqslant \mathrm{I}(\rho \neq 0)/(2|\rho|)$ *and* $\left|\partial^2 D_{ii}(\rho)/\partial\rho^2\right| = |2\rho\sum_{j=1}^n w_{ji}^2(1-3\rho^2\sum_{j=1}^n w_{ji}^2)|/(1+\rho^2\sum_{j=1}^n w_{ji})^3 \leqslant 2|\rho|\sum_{j=1}^n w_{ji}^2(1+3\rho^2\sum_{j=1}^n w_{ji}^2) \leqslant 2|\rho|C_2(1+3\rho^2C_2)$, *because* $\sum_{j=1}^n w_{ji}^2 = (\mathbf{W}^{\mathrm{T}}\mathbf{W})_{ii} \leqslant \lambda_{\max}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right) \leqslant C_2$.

Condition (C1) ensures the existence of some moments, which is very mild. Conditions (C2) and (C3) are standard assumptions used in functional linear regression and similar assumption can be found in Hall and Horowitz (2007). We can interpret Condition (C3) as a "smoothness class" of functions, where functions become smoother as τ increases. The assumption $\tau > \alpha/2+1$ can be interpreted as requiring that $\beta(t) \equiv \sum_{k=1}^{\infty} b_k\phi_k(t)$ be sufficiently smooth relative to $G(s,t)$, the smoothness of which can be implied by Condition (C2). Condition (C4) generally holds for smooth functions that are defined on finite domains. See Assumption 3 in Chen et al. (2022) for an analogous condition. Condition (C5) limits the number of the truncated terms under study. Condition (C6) is a standard assumption imposed on the score function. Condition (C7) is also a common assumption on the involved terms. See Kato (2012) for an analogous condition in functional regression. Condition (C8) requires several matrices to be nonsingular and/or to have bounded maximum eigenvalues. This is a reasonable condition, and similar conditions are widely assumed in the literature (Fan and Peng, 2004; Wang et al., 2009). These conditions are used to bound the difference between the derivative of the score function based on the estimated PFC scores and that based on true scores, as well as to facilitate the application of the central limit theorem. Condition (C9) allows certain sparsity. It only requires that each node is connected to at least $\sqrt{n}$ nodes, hence the network can be somewhat sparse, although not extremely sparse. For example, it excludes the situation where a node is only connected to a fixed number of other nodes regardless of the network size. Similar conditions can be found in Zhu et al.

(2017); Huang et al. (2019). Condition (C10) imposes constraints on the size of the non-zero coefficients and the value of the diverging constant C_n^*. See Wang et al. (2009) and Fan and Tang (2013) for discussions on these constraints.

Theorem 1. *Suppose that Conditions (C1)-(C4) hold. If we further assume that* $K \asymp n^{1/(\alpha+2\tau)}$, *then*

$$\|\widehat{\beta}_\rho(t) - \beta(t)\|^2 \quad = \quad O_p(n^{-(2\tau-1)/(\alpha+2\tau)}),$$

where $\widehat{\beta}_\rho(t)$ *denotes the estimator of functional regression coefficient at an arbitrary* ρ.

Theorem 1 rigorously establishes the convergence rate of functional regression coefficient. It demonstrates that our proposed estimator can achieve the same minimax rate $O_p(n^{-(2\tau-1)/(\alpha+2\tau)})$, which was developed in Hall and Horowitz (2007) for independent observations without network structure. The proof of Theorem 1 is provided in Supplementary Material S3.4.

Next, we show the asymptotic normality of network autoregression coefficient. To ease the presentation, we introduce some notations below. Let ρ_0 denote the true network autocorrelation coefficient, and

$$\begin{aligned}\mathscr{B}_k(\mathbf{A}) \quad = \quad & \frac{1}{n}\mathbf{b}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{-\mathrm{T}}(\rho_0)\mathbf{W}^{\mathrm{T}}\mathbf{S}(\rho_0)\mathbf{D}(\rho_0)^2\mathbf{S}(\rho_0)^{\mathrm{T}}\mathbf{A}\mathbf{e}_k \\ & + \frac{1}{n}\frac{\partial \mathbf{b}(\rho_0)^{\mathrm{T}}}{\partial \rho}\mathbf{A}^{\mathrm{T}}\mathbf{S}(\rho_0)\mathbf{D}(\rho_0)^2\mathbf{S}(\rho_0)^{\mathrm{T}}\mathbf{A}\mathbf{e}_k,\end{aligned}$$

where

$$\frac{\partial \mathbf{b}(\rho_0)^{\mathrm{T}}}{\partial \rho} = \left[\begin{array}{ccc} \frac{\partial b_1(\rho)}{\partial \rho}|_{\rho=\rho_0} & \cdots & \frac{\partial b_K(\rho)}{\partial \rho}|_{\rho=\rho_0} \end{array} \right],$$

and the form of $\partial b_k(\rho)/\partial\rho$ for $k = 1, \ldots, K$ is given in (2.13). Let

$$\begin{aligned} \varphi_{ik} = {} & \lambda_k^{-1} \int \phi_k(t) \left[Y_i^* Z_i(t) - E\left\{ Y_i^* Z_i(t) \right\} \right] dt + \lambda_k^{-1} \int E\left\{ Y_i^* Z_i(t) \right\} \\ & \times \left\{ \sum_{\ell:\ell\neq k} (\lambda_k - \lambda_\ell)^{-1} \phi_k(t) \iint \left[Z_i(u) Z_i(v) - E\left\{ Z_i(u) Z_i(v) \right\} \right] \phi_k(u) \phi_\ell(v) du dv \right\} dt \\ & - \lambda_k^{-2} (\xi_{ik}^2 - E\xi_{ik}^2) \int \phi_k(t) g(t) dt. \end{aligned}$$

Theorem 2. *Suppose that Conditions (C1)-(C8) hold and* $E(\xi_k^2 - \lambda_k)^2 < C\lambda_k^2$. *If* $\alpha > 2$, *then*

$$n^{1/4}(\widehat{\rho} - \rho_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma)$$

in distribution for the given $\mathbf{A}$, *where* $\mathbf{A} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times K}$ *and* $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^{\mathrm{T}}$ *are defined in model* (2.4), *and*

$$\Sigma = c_*^{-2} \left(n^{-1/2} E \left[\left\{ \sum_{k=1}^{K} \mathscr{B}_k(\mathbf{A}) \varphi_{jk} \right\}^2 \right] \right),$$

where c_* *is defined in* (S.67) *in Supplementary Material S3.10.*

The proof of Theorem 2 can be found in Supplementary Material S3.10. According to Theorem 2, we can draw the following conclusions.

Corollary 1. *Under the conditions in Theorem 2, when the score vector* $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^{\mathrm{T}}$

is normally distributed, we have

$$n^{1/4}(\widehat{\rho}-\rho_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\Sigma).$$

The similar normal assumption used in Corollary 1 is also adopted by Yao et al. (2005); Zhou et al. (2008); Hu and Yao (2022). It is also worth noting that the $n^{1/4}$-asymptotic normality derived in Corollary 1 is slower than the parametric rate $n^{1/2}$. Indeed, we tackle a more difficult functional regression problem, where extra complications arise from the intrinsically infinite-dimensional FPCA scores and their estimated version, which can be viewed as the true FPCA contaminated with measurement errors. More specifically, in the linear expansion form of $a_n(\widehat{\rho}-\rho_0)$, the dimensionality reduction coefficients $\widehat{b}_k(\rho)-b_k(\rho)$ are also involved, as shown in (S.27) of the Supplementary Material. To ensure that the variance of the linear expansion of $a_n(\widehat{\rho}-\rho_0)$ remains finite, we have to multiply an additional factor of $n^{-1/4}$ for the terms involving $\widehat{b}_k(\rho)-b_k(\rho)$. As a result, a_n is set to $n^{1/4}$ rather than the classical $n^{1/2}$. Whether the rate can be further improved is unclear and requires further investigation.

Next, we demonstrate that the criterion $\mathrm{BIC}^*(K)$ can consistently identify the true model. Without loss of generality, we assume $b_k \neq 0$ when $1 \leqslant k \leqslant K_0$ and $b_k = 0$ when $k > K_0$, i.e., the true model includes only the first K_0 scores as predictors. Here, K_0 is allowed to be either fixed or diverging to infinity as $n \to \infty$. We refer to the model that includes all K scores as predictors as the full model. We use integer K^* to emphasize the size of an arbitrary candidate model hereafter. The estimated scores $\widehat{\mathbf{A}}_{K*} = (\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times K^*}$ serve as the predictor variables, and the corresponding regression coefficients are $\mathbf{b}_{K*} = (b_1, \ldots, b_{K*})^{\mathrm{T}}$. We define the estimated residual variance as $\widehat{\sigma}^2_{K*} = \mathrm{SSE}_{K*}/n = \inf_{\mathbf{b}_{K*}} \|\widehat{\mathbf{S}}\mathbf{Y} - \widehat{\mathbf{A}}_{K*}\widehat{\mathbf{b}}_{K*}\|^2/n$.

Therefore, $\mathbf{A}_0$ and $\mathbf{A}_K$, along with their estimated counterparts, represent the score matrices under the true model and the full model, respectively. Throughout the article, we refer to a model as underfitted if $K^* < K_0$, and as overfitted if $K^* > K_0$. In the overfitted case, we define the set of redundant predictors $\widehat{\mathbf{A}}_0^c$ and $\mathbf{A}_0^c$, corresponding to the column index set $\{K_0+1, \ldots, K^*\}$. We also define $\widehat{\mathbf{A}}_0^{c_K}$ and $\mathbf{A}_0^{c_K}$, corresponding to the column index set $\{K_0+1, \ldots, K\}$. Under the overfitting scenario, for $k = K_0+1, \ldots, K^*$, we further define

$$\widehat{\mathbb{A}}_k^{\text{proj}} = \widehat{\mathbf{Q}}_0 \widehat{\mathbb{A}}_k \quad \text{and} \quad \mathbb{A}_k^{\text{proj}} = \mathbf{Q}_0 \mathbb{A}_k \tag{3.20}$$

where the projection matrices are given by $\widehat{\mathbf{Q}}_0 = \mathbf{I}_n - \widehat{\mathbf{A}}_0(\widehat{\mathbf{A}}_0^{\mathrm{T}}\widehat{\mathbf{A}}_0)^{-1}\widehat{\mathbf{A}}_0^{\mathrm{T}}$ and $\mathbf{Q}_0 = \mathbf{I}_n - \mathbf{A}_0(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0)^{-1}\mathbf{A}_0^{\mathrm{T}}$. In the above, the kth column of $\widehat{\mathbf{A}}_{K*}$ is denoted by $\widehat{\mathbb{A}}_k = (\widehat{\xi}_{1k}, \ldots, \widehat{\xi}_{nk})^{\mathrm{T}} \in \mathbb{R}^n$, and similarly, $\mathbb{A}_k$ denotes the kth column of the true score matrix $\mathbf{A}_{K^*}$. Note that when $K^* = K$, $\widehat{\mathbf{A}}_{K*}$ and $\mathbf{A}_{K*}$ are naturally replaced by $\widehat{\mathbf{A}}_K$ and $\mathbf{A}_K$, respectively.

For simplicity, we assume that the error term ε follows a normal distribution in the following analysis. This assumption is made solely to simplify the technical proofs and can be relaxed (Huang et al., 2008; Wang et al., 2009).

Theorem 3. *Suppose that Conditions (C1), (C2), (C5), (C8) and (C10) hold, and that the error term ε is normally distributed, we have*

$$P\left\{\min_{K^*<K_0}(BIC_{K*}) > BIC_K\right\} \to 1.$$

By Theorem 3 we know that the minimal BIC value that is associated with underfitted models is guaranteed to be larger than that of the full model as long as the sample size is sufficiently large.

Theorem 4. *Suppose that Conditions (C1), (C2), (C5), (C8) and (C10) hold, and that the error term ε is normally distributed, we have*

$$P\left\{\min_{K*>K_0}(BIC_{K*}) > BIC_{K_0}\right\} \to 1.$$

By Theorem 4, we know that, with probability tending to 1, any overfitted model cannot be selected by the BIC either, because its BIC value is not as favourable as that of the true model. Hence, Theorems 3 and 4 show that the modified BIC can identify the true model consistently.

4. Simulation Study

4.1 Preliminary setup

To evaluate finite sample performance of the proposed methods, we consider the following three simulation settings. The main difference lies in the generating mechanism of the network structure $\mathbf{W}$ (i.e., the row normalzized adjacency matrix $\boldsymbol{\mathcal{A}}$).

Scenario 1. *(Dyad Independence Model) By Holland and Leinhardt (1981), a dyad is defined as $D_{ij} = (a_{ij}, a_{ji})$ for any $1 \leqslant i < j \leqslant n$. Assume D_{ij}'s are mutually independent of each other. To allow for sparsity of the network, we define $P\{D_{ij} = (1,1)\} = 0.5n^{-1}$ and $P\{D_{ij} = (1,0)\} = P\{D_{ij} = (0,1)\} = 5n^{-1}$. Then we have $P\{D_{ij} = (0,0)\} = 1 - 5.5n^{-1}$, which is very close to 1 for a large n.*

Scenario 2. *(Stochastic Block Model) In this scenario, we consider the network structure generated from the stochastic block model (Wang and Wong, 1987; Nowicki and Snijders, 2001), which is one of the most popular network structures. Let $S = 20$ be the total*

number of blocks. First, we randomly assign each node a block label s $(s = 1, \cdots, S)$ *with equal probability* $1/S$. *Next, let* $P(a_{ij} = 1) = 0.9n^{-1}$ *if* i *and* j *belong to the same block;* $P(a_{ij} = 1) = 0.3n^{-1}$ *otherwise. Thus, nodes within the same block are more likely to be connected with each other.*

Scenario 3. *(Power-Law Distribution Network) In a social network, it is commonly observed that the majority of nodes have few links but a small proportion (i.e., celebrities) possess a large number of links Barabási and Albert (1999). The number of links usually follows the power-law distribution (Clauset et al., 2009). To mimic this phenomenon, we simulate the adjacency matrix* $\boldsymbol{\mathcal{A}}$ *according to Clauset et al. (2009) as follows. First, we generate the in-degree* $m_i = \sum_j a_{ji}$ *for node* i *by the discrete power-law distribution, i.e.,* $P(m_i = k) = ck^{-\alpha}$, *where* c *is the normalizing constant and exponent parameter* $\alpha = 2.5$. *Then, for the* i*th node,* m_i *nodes are randomly selected as its followers.*

Under each scenario, the response $\mathbf{Y} = (Y_i, i = 1, \ldots, n)$ is generated from the model

$$\begin{aligned} \mathbf{Y} &= (\mathbf{I} - \rho\mathbf{W})^{-1}\left[\left\{\int Z_i(t)\beta(t)dt, i = 1, \ldots, n\right\} + \boldsymbol{\varepsilon}\right] \\ &= (\mathbf{I} - \rho\mathbf{W})^{-1}\left[\left(\sum_{k=1}^{p_0} \xi_{ik} b_k, i = 1, \ldots, n\right) + \boldsymbol{\varepsilon}\right], \end{aligned}$$

where p_0 denotes the true number of principal components, and we consider different settings of p_0. Similar consideration can also be found in Xu et al. (2018). For $i = 1, \ldots, n$, the functional predictor $Z_i(t) = \sum_{k=1}^{p_0} \xi_{ik}\phi_k(t)$, where the corresponding kth eigenfunction $\phi_k(t)$ is the kth Fourier basis functions on $[0, 10]$. The scores $\{\xi_{ik}, i = 1, \ldots, n\}$ are i.i.d. distributed as $\mathcal{N}(0, 16k^{-2})$ for $k = 1, \ldots, p_0$. The underlying functional regression coefficient is $\beta(t) = \sum_{k=1}^{p_0} b_k\phi_k(t)$, a linear combination of the eigenfunction, where $b_1 = 0.3$ and $b_j = 4(-1)^{j+1}j^{-2}$

for $j > 1$. The corresponding network autoregression coefficient ρ is set to be 0.1 or 0.3. We further remark that although the proposed method is motivated by the assumption that random eror follows normal distribution, the method is still valid for the non-normal case, so we consider two different cases, where ε_i is generated independently from (1) $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1,\ldots,n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 = 0.01$; and (2) t-distribution with degree 5.

The above three different scenarios of $\mathbf{W}$ are typical network models often assumed in the literature, see, for example, Lei and Rinaldo (2015); Zhu et al. (2022) and references therein. The corresponding functional part follows the standard functional data setting and similar generating mechanism can be found in Kato (2012); Kong et al. (2016); Wong et al. (2019).

To gauge a reliable evaluation of the simulation study, we repeat the experiments for $M = 100$ times. The network size is set to be $n = 200$ and 500, respectively. In each simulation replicate, we generate n predictor trajectories and the observations are made at 200 equally spaced points on $[0, 10]$. To measure the estimation accuracy of the functional coefficients, we consider the integrated mean squared error (IMSE) for the predictor, which is defined as $\int_0^{10} \{\widehat{\beta}(t) - \beta(t)\}^2 dt$, where $\widehat{\beta}(t)$ is the average estimator of the M repetitions and the integral is estimated by discretizing the interval $[0, 10]$ into 20 equally spaced subintervals. We also use supremum norm (SUP) to evaluate the performance of the functional estimators, that is, $\max_{1 \leqslant k \leqslant 20} \{\widehat{\beta}(t_k) - \beta(t_k)\}^2$, where we also discretize the interval $[0, 10]$ into 20 equally spaced intervals to facilitate the analysis. For the mth replicate, write $\widehat{\rho}^{(m)}$ as the estimate of ρ. Then the bias is evaluated as Bias $= \rho - \overline{\rho}$, where $\overline{\rho} = M^{-1} \sum_{m=1}^{M} \widehat{\rho}^{(m)}$, and we also report the standard deviation of $\widehat{\rho}$, which is calculated by SD $= \{M^{-1} \sum_{m=1}^{M} (\widehat{\rho}^{(m)} - \overline{\rho})^2\}^{1/2}$. Finally, we examine the empirical performance of the proposed AIC and BIC criteria in determining

K, and provide the graphical summary for functional regression coefficient. For simplicity, we only present the simulation results for the case $\rho = 0.1$ below, while the results for $\rho = 0.3$ are provided in Supplementary Material S4.

4.2 Simulation results

We first evaluate the estimation accuracy of the functional regression coefficients and network autocorrelation coefficient for Scenario I-III. We allow the true number of principal components to vary from being fixed to diverging. Tables 1–2 contain IMSE and SUP of functional regression coefficient and Bias and SD of network autocorrelation coefficient, respectively. Both tables show that there is a general tendency for IMSE, SUP, Bias ans SD to decrease as sample size n increases. Interestingly, it is observed that the error distribution affects the performance of the functional regression coefficient, while it hardly affects the estimation of network coefficient. For instance, in Table 1, the estimated functional regression coefficient under t distribution dose not perform as well as that in the normal error case, while in Table 2, the corresponding network autocorrelation coefficient is unaffected by error distribution. In some cases, t distribution even leads to smaller Bias and SD compared to the normal distribution. This illustrates the robustness of our proposal. Fewer number of true principal components lead to better performance of functional regression coefficient. This is because there are fewer parameters to be estimated under fewer principal components. From Table 2, we also find that the estimation of network autocorrelation coefficient is fairly stable regardless of the number of principal components.

The simulation results for AIC and BIC in determining K under various settings, including the normal and t distribution cases, are summarized in Tables 3 and 4, respectively, with

the true number of principal components set to 50. Each table shows the average numbers of K selected by the four criteria, as well as the corresponding IMSE. In general, the BIC* or BIC** lead to the smallest IMSE. As expected, the performance of all IMSE improves as the sample size n increases. In the normal distribution case, in Table 3, BIC* works better than the other three criteria in most cases. On the other hand, in the t-distribution case, in Table 4, BIC** works better than others. Looking closely at these tables, we find that BIC* and BIC** perform quite stably in every case, even if they do not dominate the others in some cases. For example, in Table 3, under Scenario III and when sample size n is 500, AIC performs the best, followed by BIC* and BIC**. Similar observations apply to other cases. Thus, we recommend using BIC* and BIC** in practice.

We also provide graphical summaries for the estimated functional regression coefficient in Figure 3, where we compare in each panel the 2.5% and 97.5% pointwise percentiles of our estimators with the truth for normal and t distribution with $n = 200$, respectively. Here the number of principal components is chosen by BIC*. It is remarkable that the newly proposed estimator performs very well. The estimated coefficients are close to the true values, and the true coefficients are nicely covered by the 95% confidence bands. This is consistent with the results in the corresponding tables. As we can see, even though the true error distribution is non-Gaussian, averaged estimated curves in all panels are remarkably close to the true reference lines, which also shows that the proposed method is quite robust against mild violation of the Gaussian assumption. These results indicate the advantages of the proposed methods.

5. Empirical Application

5.1 Crop and weather data

We first illustrate the proposed procedure via empirical analysis of a crop yield data set (Park et al., 2023). The data set was collected from the National Agricultural Statistics Agency (`https://quickstats.nass.usda.gov/`), and consists of several yield-related variables at the county level (such as annual crop yield in bushels per acre, size of harvested land and the proportion irrigated land to the total harvested land) from a total of 403 counties in five Midwestern states of the United States from 1999 to 2020. In addition, some meteorology measurements for each county are also available from the National Climatic Data Center (`https://www.ncdc.noaa.gov/data-acc`), including daily maximum temperature and daily minimum temperature. The main goal of this study is to investigate the impact of daily temperature difference on corn yield across different counties in the five Midwestern states while taking into account the geographical network information.

To apply our model, we consider year 2012 which contains $n = 324$ counties. Let Y_i be the average corn yield per acre for the ith county and the functional predictor $Z_i(t)$ be the difference of daily maximum and minimum temperature trajectories in the time domain $\mathcal{I} = [0, 365]$ for $i = 1, \ldots, n$. To accommodate spatial correlation of corn yields and analyze the geographical network structure, we regard each county as a node. The adjacency matrix $\boldsymbol{\mathcal{A}}$ is naturally defined using spatial distances between any two counties. Specifically, let $\mathbf{s}_1, \ldots, \mathbf{s}_n$ be the locations of the n counties (i.e., the longitude and latitude), where $\mathbf{s}_i \in \mathbb{R}^2$. Then a_{ij} is defined as $a_{ij} = 1/\|\mathbf{s}_i - \mathbf{s}_j\|$ for $i \neq j$ and $a_{ii=0}$ for $i = 1, \ldots, n$.

The estimated functional regression coefficient together with their 95% bootstrap point-wise confidence bands is plotted in Figure 1. We observe that the overall impact of the

temperature difference on corn yield is fluctuating and can be both positive and negative. In fact, the impact fluctuates between positive and nagative in the first 100 days and between days 230 and 300, while stay negative in other days. More specifically, the maximum positive impact is at around days 90 and 280, and the maximum negative impact is around day 180, indicating that the temperature difference respectively has the strongest positive and negative correlations with corn yield at these days. We suspect that this is because July is in the hottest season of the year, and excessively high temperatures hence large temperature differences have a negative impact on crops. On the other hand, days 90 and 280 are in spring and autumn, and warm day temperatures which lead to large temperature differences in these seasons are more suitable for the growth of corn. In this case study, we focus solely on the dynamic association between the corn yield and the temperature difference trajectories as an initial attempt, because temperature difference is one of the most important factors affecting corn growth. The estimated network effect $\widehat{\rho} = 0.8775$ is positive, indicating that the corn yield of a node is positively related to its connected neighbors. This result is consistent with common knowledge as these regions generally share similar climatic and soil conditions, leading to similar yields.

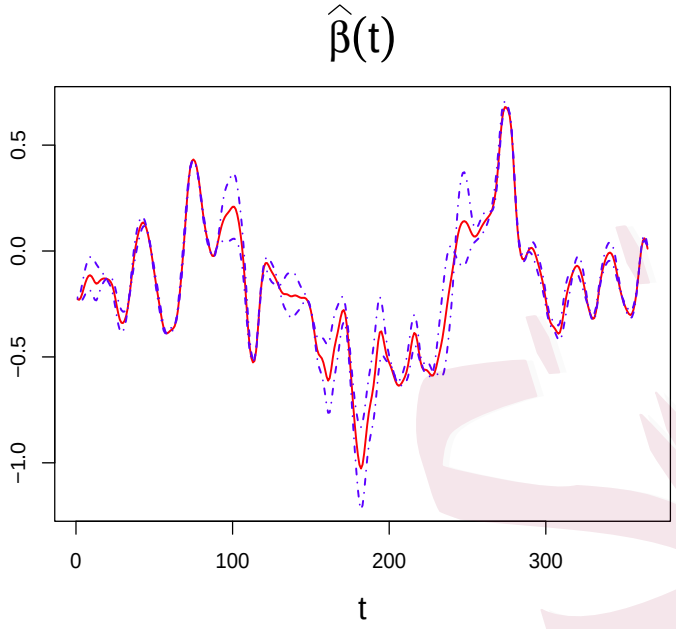

Figure 1: Estimated functional regression function (solid) and 95% point-wise confidence bands (dashed) based on 200 bootstrap samples for temperature difference.

5.2 Air pollution analysis

We further apply the proposed methodology to a large-scale air pollution dataset in the United Kingdom (Lee et al., 2017). The dataset consists of predictive data for PM2.5, NO_2, O_3 and PM10, collected from air quality monitoring stations across 1516 prediction sites within England. In addition, the dataset also includes geographic information of each site, i.e., longitude and latitude. Other studies (e.g., Fan et al., 2021) have shown that NO_2 has a larger effect on PM2.5. Therefore, in our analysis, we use the monthly concentration of NO_2 from January 2007 to December 2011 (60 months) as the functional predictor. The response of interest is the concentration of PM2.5. We investigate the dynamic influence of NO_2 on the average concentration of PM2.5 over the period of 60 months, taking into account the geographic information.

Next, we apply the newly proposed procedure to this dataset. The adjacency matrix is constructed between stations by taking advantage of their spatial locations. The solid curves in Figure 2 are the fitted regression curves along with 95% point-wise confidence bands by the proposed procedure. Overall, the concentrations of NO_2 have positive influences on that of PM2.5 during the period of 60 months. As time progresses, the impact of NO_2 on PM2.5 gradually weakens, but there are some fluctuations. Specifically, at $t = 0$, $\widehat{\beta}(t)$ is approximately 0.15, and by $t = 60$, it decreases to around 0.08. We suspect the reason of a nonconstant effect over time is due to the complex reaction between nitrogen dioxide and other elements in the air, as well as various dynamic weather conditions that are not taken into account in our model. To understand the entire picture of the air pollution issue as a dynamic system of course requires much more complex analysis that is beyond the current work. Additionally, we note that the estimated functional regression coefficient is not very smooth due to high noise level of concentrations trajectory of NO_2. The estimated network autocorrelation coefficient $\widehat{\rho} = 0.8891$ indicates a strong positive correlation between the PM2.5 concentration at each node and the surrounding PM2.5 concentration. This is because the air flow leads to similarity in PM2.5 concentration in the surrounding areas.

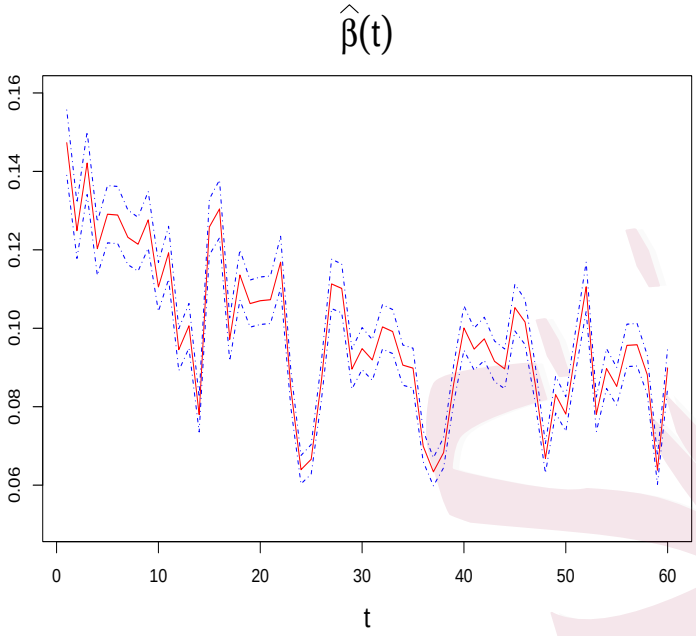

Figure 2: Plot of estimated functional regression coefficient in air pollution analysis. The solid red curve represents the fitted coefficient, while the dotted blue curve represents the corresponding 95% point-wise confidence bands.

6. Conclusion and Discussion

Compared to the classical FLR, the network FLR introduced in this article provides a more flexible description of the relationship between a scalar response and a functional predictor. The estimation procedure consists of two important steps. First, FPCA is employed to estimate FPC scores and eigenvalues of the covariance function of the functional predictor. Second, we devise a least-squares type estimator that overcomes the complexity caused by the network interdependence. Although the proposed methods are partially motivated by application with specific network, they are widely applicable in many other situations, including large-scale social network, such as the network in Facebook, Twitter, Sina Weibo, and many others.

Our work leads to several potential further research problems. On the theory side,

it would be interesting to investigate whether one can improve the $n^{1/4}$-rate of network autocorrelation coefficient estimation in Corollary 1 to achieve the standard parametric rate $n^{1/2}$. On the methodology side, it would be of great interest to develop algorithms for more complicated network functional model settings such as functional generalized linear model, functional index model, functional additive model, etc.

Another extension concerns functional responses $Y_i(t)$ with functional network effect $\rho(t)$, as suggested by an anonymous reviewer. For instance, considering the model

$$Y_i(t) = \int Z_i(s)\beta(s,t)ds + \rho(t)\sum_{j=1}^{n} w_{ij}Y_j(t) + \varepsilon_i(t), \quad i = 1, \ldots, n. \tag{6.21}$$

Building on the method proposed in this article, two possibilities arise in handling (6.21). First, when subjects are observed on regular grids, say $\{t_1, \ldots, t_M\}$, (6.21) reduces to our model (2.2) at each grid point t_m. We can apply the proposed estimation method at t_m to obtain $\widehat{\beta}(s, t_m)$ and then use smoothing or interpolation to estimate $\widehat{\beta}(s,t)$ for any t. Second, when subjects are observed at irregular time points, to overcome the hurdle of different time points for different subjects, we may need to conduct FPCA to achieve the dimension reduction for the functional predictor. In particular, we can write $Z_i(s) = \sum_k \xi_{ik}\phi_k(s)$ and $\beta(s,t) = \sum_k \beta_k(t)\phi_k(s)$, allowing (6.21) to be rewritten as

$$Y_i(t) = \sum_k \xi_{ik}\beta_k(t) + \rho(t)\sum_{j=1}^{n} w_{ij}Y_j(t) + \varepsilon_i(t), \quad i = 1, \ldots, n. \tag{6.22}$$

Note that (6.22) is structurally similar to model (1) of Zhu et al. (2022), except that the predictors $\xi_{ik}, i = 1, \cdots, n, k = 1, 2, \cdots$ need to be truncated and estimated. Based on (6.22), we then can apply kernel based least squares estimate to obtain the $\widehat{\beta}_k(t)$ (Zhu et al., 2022).

Although methodologically it is achievable, extending the model studied in this article to (6.21) that allows functional response and functional network effect can be very challenging in terms of theoretical investigation.

Supplementary Materials

The Supplementary Material contains auxiliary lemmas, additional simulation results, theoretical analysis, and all proofs of Theorems 1-4.

Acknowledgements

We are grateful to the Editor, an anonymous Associate Editor and reviewers for their careful reading and valuable suggestions, which have helped to improve the article. This research was supported by the National Key R&D Program of China under Grant 2024YFA1012200, the National Natural Science Foundation of China under Grant 12571285, and the National Institute of Health.

References

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216.

Chen, C., Guo, S., and Qiao, X. (2022). Functional linear regression: dependence and error contamination. *Journal of Business & Economic Statistics*, 40(1):444–457.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37(1):35–72.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fan, Y., Liu, Y., and Zhu, L. (2021). Optimal subsampling for linear quantile regression models. *Canadian Journal of Statistics*, 49(4):1039–1057.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(3):531–552.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Hu, X. and Yao, F. (2022). Sparse functional principal component analysis in high dimensions. *Statistica Sinica*, 32(4):1939–1960.

Huang, D., Lan, W., Zhang, H. H., and Wang, H. (2019). Least squares estimation of spatial autoregressive models for large-scale social networks. *Electronic Journal of Statistics*, 13(1):1935–7524.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618.

Kato, K. (2012). Estimation in functional linear quantile regression. *The Annals of Statistics*, 40(6):3108–3136.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Boca Raton: CRC Press.

Kong, D. H., Xue, K. J., Yao, F., and Zhang, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika*, 103(1):147–159.

Le, C. M. and Li, T. (2022). Linear regression and its inference on noisy network-linked data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1851–1885.

Lee, D., Mukhopadhyay, S., Rushworth, A., and Sahu, S. K. (2017). A rigorous statisti-

cal framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. *Biostatistics*, 18(2):370–385.

Lee, L.-F., Liu, X., and Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13(2):145–176.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237.

Li, Y. H., Wang, N., and Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association*, 105(490):621–633.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Park, Y., Li, B., and Li, Y. (2023). Crop yield prediction using bayesian spatially varying coefficient models with functional predictors. *Journal of the American Statistical Association*, 118(541):70–83.

Ren, Y., Li, Z., Zhu, X., Gao, Y., and Wang, H. (2024). Distributed estimation and inference for spatial autoregression model with large scale networks. *Journal of Econometrics*, 238(2):105629.

Tho, Z. Y., Ding, D., Hui, F. K., Welsh, A., and Zou, T. (2023). On the robust estimation of spatial autoregressive models. *Econometrics and Statistics, https://doi.org/10.1016/j.ecosta.2023.01.004*.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(3):671–683.

Wang, X., Zhu, H., and Initiative, A. D. N. (2017). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112(519):1156–1168.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19.

Wong, R. K., Li, Y., and Zhu, Z. (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 114(525):406–418.

Xu, Y., Li, Y., and Nettleton, D. (2018). Nested hierarchical functional data modeling and inference for the analysis of functional plant phenotypes. *Journal of the American Statistical Association*, 113(522):593–606.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903.

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95(3):601–619.

Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474):630–643.

Zhu, X., Cai, Z., and Ma, Y. (2022). Network functional varying coefficient model. *Journal of the American Statistical Association*, 117(540):2074–2085.

Zhu, X. N., Huang, D. Y., Pan, R., and Wang, H. S. (2020). Multivariate spatial autoregressive model for large scale social networks. *Journal of Econometrics*, 215(2):591–606.

Zhu, X. N., Pan, R., Li, G. D., Liu, Y. W., and Wang, H. S. (2017). Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123.

School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, China

E-mail: yan@jsnu.edu.cn

Department of Statistics, The Pennsylvania State University, State College, Pennsylvania, USA

E-mail: yzm63@psu.edu

Table 1: Simulation results for $n = 200, 500$ with 100 replicates of the dyad independence model, stochastic block model and power-law distribution model respectively. The results are displayed when the error follows normal distribution and t-distribution in each scenario. The IMSE and SUP are both reported. The true number of principal components (# PC) ranges from 2 to 50, with $\rho = 0.1$.

		# PC: 2		# PC: 10		# PC: 50	
Scenario	n	IMSE	SUP	IMSE	SUP	IMSE	SUP
Case 1: Normal distribution							
I	200	6.79×10^{-6}	1.83×10^{-5}	6.89×10^{-4}	3.16×10^{-4}	1.02×10^{-2}	7.01×10^{-2}
	500	1.63×10^{-6}	4.82×10^{-6}	2.63×10^{-4}	1.05×10^{-3}	1.87×10^{-3}	1.30×10^{-2}
II	200	5.38×10^{-6}	1.49×10^{-5}	6.92×10^{-4}	2.92×10^{-3}	1.09×10^{-2}	7.75×10^{-2}
	500	1.09×10^{-6}	3.50×10^{-6}	2.53×10^{-4}	9.96×10^{-4}	1.88×10^{-3}	1.24×10^{-2}
III	200	4.71×10^{-6}	1.39×10^{-6}	7.78×10^{-4}	3.44×10^{-3}	0.01	0.07
	500	1.08×10^{-6}	3.52×10^{-6}	2.70×10^{-4}	1.06×10^{-3}	1.90×10^{-3}	1.31×10^{-2}
Case 2: t-distribution							
I	200	2.78×10^{-4}	5.79×10^{-4}	1.59×10^{-2}	7.15×10^{-2}	0.37	2.7
	500	1.02×10^{-4}	2.13×10^{-4}	6.55×10^{-3}	2.90×10^{-2}	0.14	0.98
II	200	3.00×10^{-4}	6.21×10^{-4}	1.59×10^{-2}	7.23×10^{-2}	0.36	2.48
	500	1.14×10^{-4}	2.38×10^{-4}	6.69×10^{-3}	2.99×10^{-2}	0.16	1.09
III	200	2.18×10^{-4}	4.61×10^{-4}	1.69×10^{-2}	7.42×10^{-2}	0.39	2.74
	500	8.42×10^{-5}	1.78×10^{-4}	6.70×10^{-4}	2.99×10^{-2}	0.14	1.03

Table 2: Simulation results for $n = 200, 500$ with 100 replicates of the dyad independence model, stochastic block model and power-law distribution model respectively. The results are displayed when the error follows normal distribution and t-distribution in each scenario. The Bias and SD are both reported. The true number of principal components (# PC) ranges from 2 to 50, with $\rho = 0.1$.

		# PC: 2		# PC: 10		# PC: 50	
Scenario	n	Bias	SD	Bias	SD	Bias	SD
Case 1: Normal distribution							
I	200	-0.0838	0.0773	-0.0951	0.0738	-0.0860	0.0767
	500	-0.0471	0.0447	-0.0546	0.0555	-0.0394	0.0485
II	200	-0.0128	0.0162	-0.0096	0.0136	-0.0100	0.0141
	500	-0.0052	0.0064	-0.0041	0.0065	-0.0034	0.0052
III	200	-0.0096	0.0256	-0.0111	0.0274	-0.0137	0.0475
	500	-0.0058	0.0103	-0.0047	0.0119	-0.0087	0.0125
Case 2: t-distribution							
I	200	-0.0287	0.1652	-0.0273	0.1415	-0.0439	0.1687
	500	-0.0244	0.0976	-0.0224	0.1057	-0.0304	0.1085
II	200	-0.0079	0.0361	-0.0073	0.0428	-0.0120	0.0372
	500	0.0013	0.0246	-0.0012	0.0249	-0.0072	0.0231
III	200	-0.0057	0.0605	-0.0105	0.0531	-0.0121	0.0453
	500	-0.0019	0.0323	-0.0045	0.0307	-0.0056	0.0315

Table 3: The results are displayed as the error follows follows normal distribution in three scenarios and the sample size n is set as 200 and 500 respectively. Average numbers of K selected by four criteria as well as the corresponding IMSE in the case of $\rho = 0.1$ are shown. We report the selected $\widehat{K}$ in the first line. We report the corresponding IMSE values in the second line (The original IMSE values multiplied by 10^3).

n	Scenario		AIC	BIC	BIC*	BIC**
200	I	$\widehat{K}$	9.46	7.59	**9.61**	8.69
		IMSE	1.66	1.87	**1.61**	1.70
	II	$\widehat{K}$	7.66	6.33	**8.00**	6.94
		IMSE	1.87	2.20	**1.76**	1.86
	III	$\widehat{K}$	8.20	6.40	**8.40**	7.43
		IMSE	2.01	2.17	**1.80**	1.82
500	I	$\widehat{K}$	12.71	9.89	**11.92**	11.09
		IMSE	0.55	0.69	**0.52**	0.58
	II	$\widehat{K}$	9.37	7.76	**9.32**	8.61
		IMSE	0.74	1.07	**0.73**	0.85
	III	$\widehat{K}$	**10.07**	8.17	9.92	9.05
		IMSE	**0.65**	0.98	0.67	0.78

Table 4: The results are displayed as the error follows t distribution in three scenarios and the sample size n is set as 200 and 500, respectively. Average numbers of K selected by four criteria as well as the corresponding IMSE in case of $\rho = 0.1$ are shown. We report the selected $\widehat{K}$ in the first line. We report the corresponding IMSE values in the second line (The original IMSE values multiplied by 10^3).

n	Scenario		AIC	BIC	BIC*	BIC**
200	I	$\widehat{K}$	5.41	3.84	6.13	**4.69**
		IMSE	2.29	0.98	1.93	**0.97**
	II	$\widehat{K}$	5.36	3.80	6.15	**4.31**
		IMSE	1.85	0.96	2.08	**0.92**
	III	$\widehat{K}$	5.59	3.87	6.68	**4.52**
		IMSE	2.72	1.06	2.32	**1.03**
500	I	$\widehat{K}$	6.69	4.44	6.76	**4.97**
		IMSE	1.82	0.62	0.85	**0.53**
	II	$\widehat{K}$	6.83	4.49	6.93	**5.05**
		IMSE	1.79	0.59	1.22	**0.54**
	III	$\widehat{K}$	6.40	4.32	6.40	**4.89**
		IMSE	2.15	0.62	0.68	**0.56**

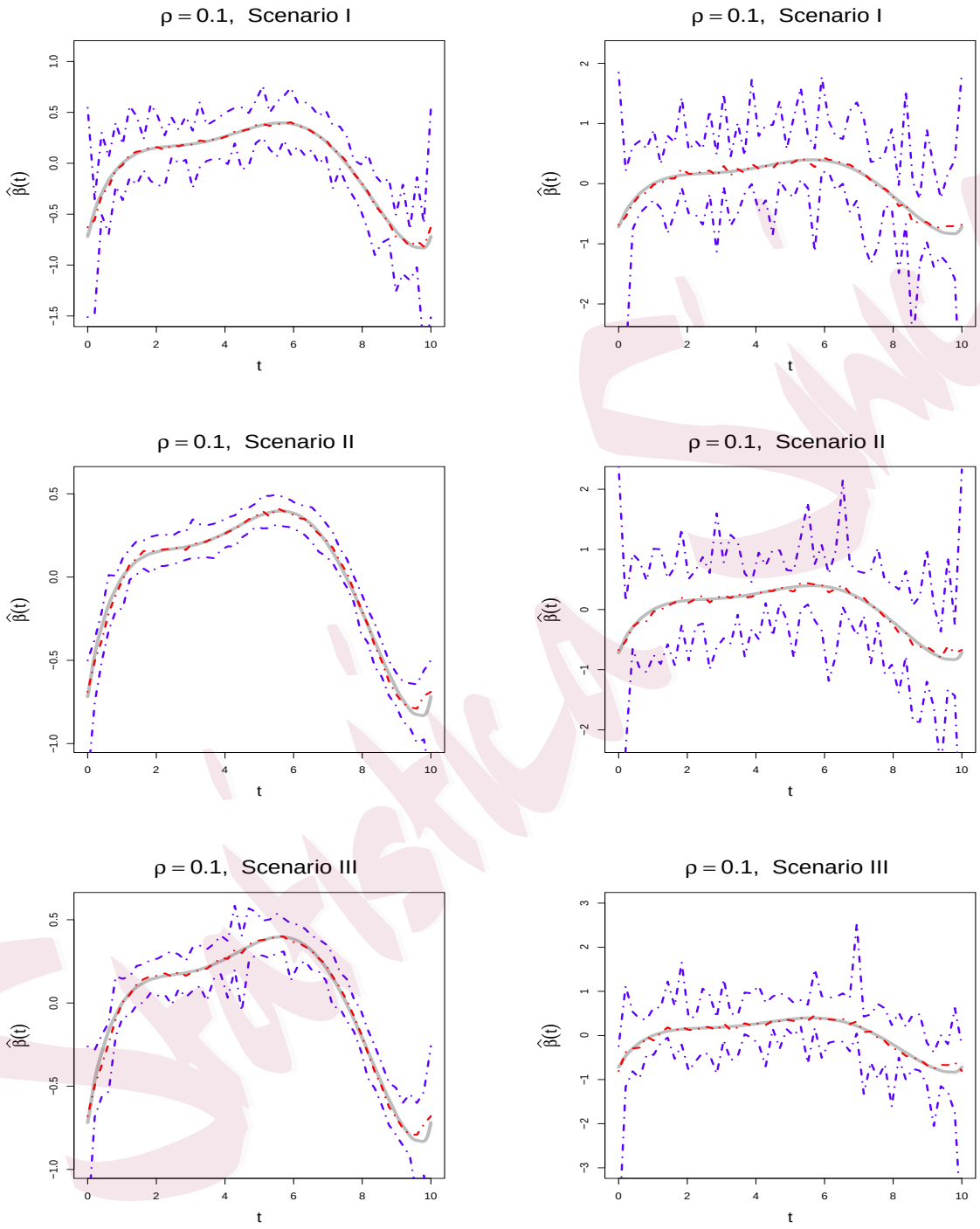

Figure 3: The functional coefficients estimation $\widehat{\beta}(t)$ under normal distribution and t distribution, respectively, with sample size $n = 200$ and $\rho = 0.1$. The left panels contain the results under normal distribution, while the right panels under t distribution. In each panel, the solid grey line is the true value, the dashed red line is the average estimated value, and the dashed blue lines are the pointwise 2.5% and 97.5% percentiles of the estimators based on 100 replications.