Spatial Functional Linear Model and its Estimation Method

Tingting Huang, Gilbert Saporta, Huiwen Wang, Shanshan Wang

Spatial Functional Linear Model and Its Estimation Method

Tingting Huang [1], Gilbert Saporta [2], Huiwen Wang [1,3], and Shanshan Wang [1,4]

[1] School of Economics and Management, Beihang University, Beijing, China

[2] CEDRIC CNAM, 292 rue St Martin, 75141 Paris Cedex 03, France

[3] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China

[4] Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations, Beijing, China

---

Address for correspondence: Shanshan Wang, School of Economics and Management, Beihang University, Xueyuan Road No.37, Haidian District, Beijing, China.
E-mail: sswang@buaa.edu.cn.
Phone: (+86) 010 82339337.
Fax: (+86) 010 82328037.

Abstract: The classical functional linear regression model (FLM) and its extensions, which are based on the assumption that all individuals are mutually independent, have been well studied and are used by many researchers. This independence assumption is sometimes violated in practice, especially when data with a network structure are collected in scientific disciplines including marketing, sociology and spatial economics. However, relatively few studies have examined the applications of FLM to data with network structures. We propose a novel spatial functional linear model (SFLM), that incorporates a spatial autoregressive parameter and a spatial weight matrix into FLM to accommodate spatial dependencies among individuals. The proposed model is relatively flexible as it takes advantage of FLM in handling high-dimensional covariates and

spatial autoregressive (SAR) model in capturing network dependencies. We develop an estimation method based on functional principal component analysis (FPCA) and maximum likelihood estimation. Simulation studies show that our method performs as well as the FPCA-based method used with FLM when no network structure is present, and outperforms the latter when network structure is present. A real weather data is also employed to demonstrate the utility of the SFLM.

Key words: FPCA; Functional linear model; Maximum likelihood estimation; Network structure; Spatial autoregressive model

1 INTRODUCTION

With the rapid development of electronic technology, huge amounts of data can be collected and stored cheaply. In particular, some types of data that are recorded at high frequencies can be regarded as almost continuously observed. Examples include trajectory data, weather data and stock data. These types of data are called functional data, which are inherently infinite-dimensional and are rich in information. Hence, functional data analysis (FDA) has been applied in many subject areas, such as biology, the medical sciences, meteorology, econometrics, finance, chemometrics and geophysics. For examples, see Ramsay and Silverman (2002) and Horváth and Kokoszka (2012).

Functional regression is one of the most important tools in FDA. Classical functional linear regression model (FLM) represents the simplest form of functional regression (an overview is given in Paganoni and Sangalli (2017)), and receives much attention in FDA. Studies related to FLM include Cai and Hall (2006); Hall and Horowitz (2007); Ferraty et al. (2013); Zhou et al. (2013); Brockhaus et al. (2015), and reference therein. FLM has become increasingly popular because they can be applied to problems that are difficult to address in the framework of scalar and vector observations.

Generally, let Y be a scalar response variable, and let $X(t)$ be a second-order stochastic

process on a compact interval Γ. Moreover, we assume that $X(t)$ is square integrable with zero mean and $E(\int_\Gamma X^2(t)dt) < \infty$. Formally, a FLM can be written as

$$Y = \alpha + \int_\Gamma X(t)\beta(t)dt + \epsilon, \tag{1.1}$$

where α is the intercept, $\beta(t)$ is the unknown slope function, and ϵ is the random error term. This term is independent of $X(t)$ and has zero mean and finite variance.

During the past few years, extensions of the FLM (1.1) have been studied to address specific problems. For example, James (2002) proposed functional logistic regression and functional censored regression to address cases in which the responses of FLM is binary and right censoring, respectively. Subsequently, Escabias et al. (2004) presented some alternative methods for estimating the parameter function in functional logistic regression model based on principal components. Aneiros-Pérez and Vieu (2006) constructed a semi-functional partial linear model that combines the advantages of semi-linear model and nonparametric statistics for functional data. Ferraty et al. (2013) generalized the FLM to functional projection pursuit regression, which allows for more interpretability. Liu et al. (2017) presented a functional linear mixed model to investigate the scalar and functional covariate effects of both individuals and population.

All of the methods mentioned above assume that all the individuals are mutually independent. However, given the rapid advances in information technology, relation information among individuals can easily be collected. Network-structured data, which represent one common form of relation information, are becoming increasingly available. In such data, the realizations of the dependent variable are correlated with each other. If we use FLM or its aforementioned variations of FLM to model this kind of data, the information contained in the data may not be fully exploited; moreover, the inferences obtained when network effects are ignored may be misleading. These observations motivate us to develop a novel statistical model for application to functional data with network structures.

To better illustrate our motivation for carrying out this study and its significance, we

show a motivating example. We collected weather data from the China Meteorological Yearbook covering the period between 2005 and 2007. These data record monthly temperatures and precipitation in 34 major cities in China. Our aim is to investigate the effect of temperature on precipitation over these three years. The scalar response is the logarithm of the mean annual precipitation, and the functional covariate is the mean monthly temperature. In a preliminary analysis, we employ the Moran's I statistic to test whether spatial autocorrelation exists among the responses. Unfortunately, the value of the Moran's I statistic is 0.7, and the P value is smaller than 0.001, which indicates that there is a significant correlation among these responses. After we apply the FLM directly to the weather data, the spatial dependence among the residuals of the FLM persists; the Moran I statistic is 0.5, and its P value is smaller than 0.001. We also display the Moran's I scatterplot of the response Y and the residuals of the FLM in Figure 1.

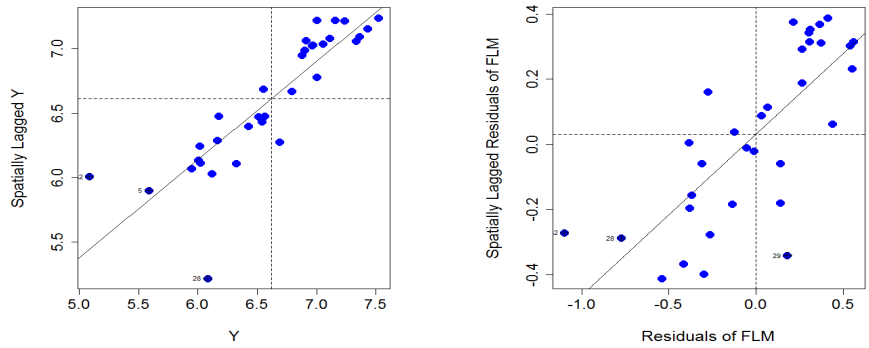

Figure 1: Moran's I scatterplot of Y (left) and the residuals of the FLM (right), respectively

Figure 1 shows that the responses and the residuals of FLM display a significance spatial autocorrelation. Thus, FLM may not be appropriate for such data. This result motivates the incorporation of spatial correlation in analyses of spatially correlated functional data. A detailed analysis of this weather data can be found in Section 5.

Fortunately, when the predictor is scalar, instead of functional, spatial autoregressive (SAR) model is frequently used model to accommodate the dependencies introduced by network structures. SAR model has been applied to social interactions, strategic

interactions and geostatistics. Studies that apply SAR model include Case (1991), Topa (2001), and Olubusoye et al. (2016), among others. Moreover, estimation methods for SAR model are described in Ord (1975), Lee (2004), Kelejian and Prucha (1999), Lee (2007) and Lesage and Pace (2009). In a SAR model, a spatial weight matrix is employed to denote adjacent relations among the observations, and an unknown spatial autoregressive parameter is used to reflect the strength of neighbouring effects. By borrowing concepts from SAR model, it is straightforward to incorporate the network structures into FLM model using a spatial autocorrelation parameter and a weight matrix. The proposed model is here named the spatial functional linear model (SFLM).

The new SFLM displays powerful capabilities in addressing the functional covariate in the model, together with the spatial dependencies between the outcomes of adjacent units. To obtain estimates of the unknown parameters, we develop an easily implemented estimation method that employs the FPCA method to handle functional data; moreover, maximum likelihood estimator for SAR model (Ord (1975), Lee (2004)) is used to handle the spatial parameters. The simulation results show that our estimation method performs well. In particular, the new method behaves as well as the FPCA-based method used with FLM when the spatial autocorrelation parameter equals zero, in which case the SFLM degenerates into an FLM; moreover, the new method displays better performance than the latter when network dependencies are present. A real dataset is employed to illustrate the practical utility of the proposed model.

We also note that Pinedaros and Giraldo (2016) proposed a model that combines a spatial error model (SEM) with an FLM to accommodate network structure in functional data. The most important difference between our model and his is that the spatial dependencies in our model are contained in the response, whereas they are contained in the disturbances in the existing model. Pinedar also used maximum likelihood estimation method to estimate the parameters of his model; however, he did not provide a practical implementation method. Moreover, our methods, which were developed for use with SFLM, can be easily extended to his functional SAR model and other spatial regression models. Note that our model handles scalar responses. This feature differs

strongly from those of Aguilera-Morillo et al. (2016) and Ramn et al. (2017), which consider functional responses in spatial functional data.

We organise this article as follows. In Section 2, we formulate the new model. The proposed estimation method is constructed based on FPCA and the MLE in Section 3. The finite-sample performance of the proposed estimators is evaluated through simulation studies and is compared with the finite-sample performance of the FLM in Section 4. Finally, in Section 5, we use real data to document the usefulness of this methodology. We conclude the article with a discussion in Section 6.

2 MODEL SPECIFICATION

Following Qu and Lee (2015), we consider the spatial processes located on an unevenly spaced lattice $D \subseteq R^d$, $d \geq 1$. We observe $\{(x_i(t), y_i)\}_{i=1}^n$ from n spatial units on D. Here, $x_i(t)$ represents independent and identically distributed (i.i.d.) samples from $X(t)$, where $X(t)$ is a square integrable second-order stochastic process defined on a compact set Γ with $E(X(t)) = 0$ and $E(\int_\Gamma X^2(t)dt) < \infty$. Without loss of generality, we presume Γ is the unit, i.e. $t \in [0, 1]$.

We denote $\boldsymbol{x}(t) = \big(x_1(t), x_2(t), \cdots, x_n(t)\big)'$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$. We then formulate the SFLM as

$$\boldsymbol{y} = \alpha\boldsymbol{\tau}_n + \rho\boldsymbol{W}\boldsymbol{y} + \int_0^1 \boldsymbol{x}(t)\beta(t)dt + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\alpha\boldsymbol{\tau}_n$ is the intercept. Here $\boldsymbol{\tau}_n$ represents an n-dimensional vector of ones; α denotes a scalar parameter; ρ is the unknown spatial autocorrelation parameter that takes values within the range of $[0, 1)$; $\boldsymbol{W} = (w_{ii'})_{n\times n}$ is a pre-specified spatial weight matrix, in which $w_{ii'}$ represents the weight between units i and i'; $\beta(t)$ is a square integrable coefficient function defined on $[0, 1]$; and ϵ is the noise term, which is independent of $\boldsymbol{x(t)}$. We assume that $\boldsymbol{\epsilon}$ follows a multivariate normal distribution with zero mean and a constant diagonal covariance matrix $\sigma^2\boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is a $n \times n$ identity

matrix.

In the model presented in (2.1), the spatial weight matrix $\boldsymbol{W}$ is exogenous. This matrix is constructed according to the distances between the units on D in different contexts. For the case of geographic locations, $\boldsymbol{W}$ is formed based on adjacency relations or the nearest k neighbours, as measured in terms of the Euclidean distance or another appropriate metric. In social networks, W is established on the basis of friend relationships among individuals. In other practical settings, such as economics, $\boldsymbol{W}$ can also be built using economic factors such as GDP. In general, the spatial matrix $\boldsymbol{W}$ is standardized to be a row-normalized matrix; in this matrix, the summation of the row elements is unity, and the entries on the diagonal are zeros.

In addition, ρ is a scale parameter that reflects the impact of the neighbours. Greater values of ρ indicate that y_i is more strongly affected by its neighbours. The SFLM clearly reduces to the classical FLM when $\rho = 0$ and the SAR model when $\boldsymbol{x(t)}$ is a constant covariate that does not depend on t. Thus, the proposed SFLM is a more flexible model, given its incorporation of both functional covariates and network structures.

To provide better insight into the new model, we divide right-hand side of equation (2) into two parts. The first part is $\rho\boldsymbol{W}\boldsymbol{y}$, which represents the neighbour effects. In real cases, such effects may arise due to competition, spillover or shared sources. The second part, $\int_0^1 \boldsymbol{x(t)}\beta(t)dt$, indicates the effects of exogenous functional variables. The outcome of unit i is affected by the outcomes of its neighbours' i' ($i' \neq i$) as well as its individual-specific covariate $x_i(t)$.

Morover, we can reformulate equation (2.1) as the following equivalent expression,

$$\boldsymbol{y} = (\boldsymbol{I}_n - \rho\boldsymbol{W})^{-1}\alpha\boldsymbol{\tau}_n + (\boldsymbol{I}_n - \rho\boldsymbol{W})^{-1}\int_0^1 \boldsymbol{x(t)}\beta(t)dt + (\boldsymbol{I}_n - \rho\boldsymbol{W})^{-1}\boldsymbol{\epsilon}, \tag{2.2}$$

which shows how y is generated. y_i is clearly influenced by its neighbours' covariates $x_{i'}(t)$s, as is $i' \neq i$. The error terms in equation (2.2) are no longer independent. Given that this dependency violates the Gauss-Markov assumption, the ordinary least-squares

(OLS) estimator is not adequate to estimate the parameters in (2.1).

3 ESTIMATION METHOD

We estimate the intercept α, the spatial autocorrelation parameter ρ, the slope function $\beta(t)$, and the variance of the error term σ^2 in the SFLM (2.1) using the maximum likelihood approach combined with a basis expansion in terms of the FPCA, as shown follows.

3.1 Basis expansion based on FPCA

Note that, before estimating the parameters of our model, data representation methods, such as smoothing and interpolation, should be used to convert discretely recorded data $x_i(t_i)$ to a curve $x_i(t)$.

Let $K(s,t)$ denote the covariance function for $X(t)$, i.e. $K(s,t) = \text{Cov}(X(t), X(s))$. By Mercer's theorem, the spectral decomposition of $K(s,t)$ is then

$$K(s,t) = \sum_{j=1}^{\infty} k_j \varphi_j(s)\varphi_j(t)$$

, where $k_1 > k_2 > \cdots > 0$ are eigenvalues and $\{\varphi_j(t)\}_{j=1}^{\infty}$ are the corresponding orthogonal eigenfunctions. According to the Karhunen-Loève expansion, $X(t)$ can be expanded as

$$X(t) = \sum_{j=1}^{\infty} a_j \varphi_j(t),$$

where the a_js are uncorrelated random variables with mean zero and variance $E(a_j^2) = k_j$ with $a_j = \int_0^1 X(t)\varphi_j(t)dt$.

For an observation $\{y_i, x_i(t)\}_{i=1}^n$, the empirical version of $K(s,t)$ is

$$\hat{K}(s,t) = \frac{1}{n}\sum_{i=1}^{n} x_i(s)x_i(t) - \bar{x}(s)\bar{x}(t),$$

where $\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$. Moreover, it can be shown that

$$\hat{K}(s,t) = \sum_{j=1}^{n} \hat{k}_j \hat{\varphi}_j(s)\hat{\varphi}_j(t),$$

where $\hat{k}_j$ and $\hat{\varphi}_j(t)$ are the estimators of $\varphi_j(t)$ and k_j, respectively. For the ith observation $x_i(t)$, the estimator of a_{ij} is then $\hat{a}_{ij} = \int_0^1 x_i(t)\hat{\varphi}_j(t)$, and $x_i(t)$ can be written as $x_i(t) = \sum_{j=1}^{\infty} \hat{a}_{ij}\hat{\varphi}_j(t)$.

Similarly, based on the estimated orthonormal functional basis $\{\hat{\varphi}_j(t)\}_{j=1}^{\infty}$, $\beta(t)$ has the expression $\beta(t) = \sum_{j=1}^{\infty} b_j\hat{\varphi}_j(t)$, with $b_j = \int_0^1 \beta(t)\hat{\varphi}_j(t)dt$. Therefore, SFLM has the equivalent expression

$$\boldsymbol{y} = \alpha\boldsymbol{\tau}_n + \rho\boldsymbol{W}\boldsymbol{y} + \sum_{j=1}^{n} \hat{\boldsymbol{a}}_{\boldsymbol{j}} b_j + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\hat{\boldsymbol{a}}_{\boldsymbol{j}} = (\hat{a}_{1j}, \hat{a}_{2j}, \cdots, \hat{a}_{nj})'$.

In reality, the first few principal components (PCs), as ranked by their eigenvalues, often provide an adequate approximation of $x_i(t)$. Here, we choose the first m PCs and ignore the truncation error. The truncated SFLM of (3.1) takes the form of

$$\boldsymbol{y} \approx \alpha\boldsymbol{\tau}_n + \rho\boldsymbol{W}\boldsymbol{y} + \sum_{j=1}^{m} \hat{\boldsymbol{a}}_{\boldsymbol{j}} b_j + \boldsymbol{\epsilon}. \tag{3.2}$$

3.2 Maximum Likelihood Estimation for the Truncated SFLM

Here we focus on (3.2) and adopt the popular MLE approach to estimate the unknown parameters.

Defining $\boldsymbol{A} = (\hat{a}_{ij})_{n\times m}$, $\boldsymbol{b} = (b_1, b_2, \cdots, b_m)'$, $\boldsymbol{Z} = (\boldsymbol{\tau}_n, \boldsymbol{A})$ and $\boldsymbol{\delta} = (\alpha, \boldsymbol{b}')'$, we can write the truncated equation (5) as

$$\boldsymbol{y} \approx \rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}. \tag{3.3}$$

Based on the assumption that the error term in (2.1) follows a multivariate normal

distribution, the log-likelihood function for $\boldsymbol{y}$ is

$$\ln L(\rho, \boldsymbol{\delta}, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) + \ln|\boldsymbol{I}_n - \rho\boldsymbol{W}| - \frac{\boldsymbol{e}'\boldsymbol{e}}{2\sigma^2}, \tag{3.4}$$

where $\boldsymbol{e} = \boldsymbol{y} - \rho\boldsymbol{W}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\delta}$. Given ρ, the maximum likelihood estimate of $\boldsymbol{\delta}$ and σ^2 is

$$\hat{\boldsymbol{\delta}}(\rho) = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{I_n} - \rho\boldsymbol{W})\boldsymbol{y}, \tag{3.5}$$

$$\hat{\sigma}^2(\rho) = \frac{1}{n}(\boldsymbol{y} - \rho\boldsymbol{W}\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\delta}})'(\boldsymbol{y} - \rho\boldsymbol{W}\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\delta}}(\rho)). \tag{3.6}$$

Substituting $\hat{\boldsymbol{\delta}}(\rho)$ and $\hat{\sigma}^2(\rho)$ into (3.4) and dropping the constant term yields the following profile log-likelihood function

$$\ln L(\rho) = -\frac{n}{2}\ln(\hat{\sigma}^2(\rho)) + \ln|\boldsymbol{I}_n - \rho\boldsymbol{W}|, \tag{3.7}$$

The optimization of the maximum log-likelihood function (3.7) then reduces to a univariate optimization problem. That is, the estimator of ρ can be obtained via the maximization of (3.7), i.e.

$$\hat{\rho} = \arg\max_{\rho} \ln L(\rho). \tag{3.8}$$

Replacing ρ with $\hat{\rho}$ in (3.5) and (3.6) yields the estimators of $\boldsymbol{\delta}$ and σ^2, respectively. The estimator $\hat{\beta}(t)$ for $\beta(t)$ can be reconstructed by

$$\hat{\beta}(t) = \sum_{j=1}^{m} \hat{b}_j \hat{\varphi}_j(t). \tag{3.9}$$

where $\hat{b}_j$ is obtained from $\hat{\boldsymbol{\delta}}$ in (3.5).

In summary, the estimators $\hat{\rho}$, $\widehat{\alpha}$ and $\widehat{\sigma}^2$ of SFLM are obtained from the MLE of the truncated SFLM; on the other hand, the estimator $\hat{\beta}(t)$ of the SFLM is constructed using the functional principal component basis presented in Section 3.1 and the coefficient estimator given in Section 3.2. For convenience, we here summarize the main steps of the estimation procedure in Algorithm 1.

Algorithm 1 Main steps of the estimation procedure

1: Represent the functional predictor and slope function using the functional principal component basis. The estimation procedure is then simplified, as the remaining process is similar to the problem of estimating a SAR model. In this step, after an appropriate truncation parameter is given, the SFLM approximates a SAR model whose covariates are the principal component scores of $x_i(t)$s, as shown in (3.3).
2: Determine the estimators of the unknown parameters of the truncated SFLM obtained in Step 1. The maximum likelihood estimation method is used to obtain the spatial autocorrelation parameter ρ in (3.8), the estimators of the coefficients b, the intercept α and the variance of the error term σ^2, as shown in (3.5)-(3.6).
3: Determine the estimator of $\beta(t)$ in the SFLM. The slope function is constructed using the FPC basis mentioned in Step 1 and the coefficients estimated in Step 2, as shown in (3.9). The other estimators are obtained directly from Step 2.

3.3 Choosing the truncation parameter for the SFLM

There are two ways to determine the truncation parameter. The first method is the percentage of variance explained (PVE) for predictors, which uses eigenvalues to choose the number of PCs. The second method is derived from a Markov chain Monte Carlo model composition methodology labeled MC^3.

To compare the new SFLM with the FLM in numerical experiments, we use the first method to determine the values of this parameter. In this case, if we choose the PVE to be 80%, the truncation parameter m is subject to $\min_l \{(\sum_{j=1}^{l} \hat{k}_j)/(\sum_{j=1}^{\infty} \hat{k}_j) \geq 80\%\}$.

The second method specifies the covariates in (3.3) according to the posterior model probability. We adopt this technique to handle the weather data in Section 5. In this Bayesian methodology, prior distributions are assigned to the parameters in the truncated SFLM. Specifically, σ^2 follows an inverse gamma distribution, $\pi(\sigma^2) \sim IG(a, b)$; $\boldsymbol{b}'$ follows a multivariate normal distribution conditional on σ^2, i.e. $\pi(\boldsymbol{b}'|\sigma^2) \sim N[\boldsymbol{b}'_0, \sigma^2(g\boldsymbol{A}'\boldsymbol{A})^{-1}]$; and ρ follows a Beta prior distribution, $\pi(\rho) \sim \frac{1}{Beta(d,d)} \frac{(1+\rho)^{d-1}(1-\rho)^{d-1}}{2^{2d-1}}$. We set $a = 0, b = 0, g = \frac{1}{n}, d = 1.01$ for use with the weather data. Note that $\boldsymbol{A}$ must be scaled during preprocessing. Details of this procedure are given in Lesage and Parent (2007).

4 SIMULATION STUDY

We conduct several simulation studies to evaluate the finite-sample performance of the proposed estimators of ρ and $\beta(t)$. All of the computations were carried out in the R environment, and we use existing functions in the R packages 'spdep', 'fda' and 'fda.usc' to implement the proposed procedure.

Because spatial networks are of interest in this study, we compare the proposed SFLM with the existing FLM in terms of the behaviour of the estimators of $\beta(t)$. In particular, the SFLM is estimated using the proposed method, whereas the FLM employs an FPCA-based estimation approach. To make these two estimation methods comparable, we set the truncation parameter of FPCA to be identical to the PVE, which equals 70%. Different degrees of spatial effects are considered; the spatial parameter ρ is set to 0, 0.5, and 0.8. Note that, when $\rho = 0$, the SFLM reduces to the FLM.

As for the spatial scenario, we adopt the rook matrix by randomly apportioning n agents on a regular square grid of cells; each agent is located on a cell. In this context, if the grid has R rows and T columns, then the sample size $n = R \times T$. Units that share an edge are neighbours. This definition ensures the units in the inner field of the grid have four neighbours, the units in the corners have two neighbours, and the units along the boarders have three neighbours. Therefore, the spatial matrix is an adjacent matrix with each entry $w_{ii'} = 1$ if units i and i' are neighbours and $w_{ii'} = 0$ otherwise. We set $n = \{10 \times 30, 20 \times 25, 30 \times 30\}$ in the simulation. The spatial weight matrix is row-normalized in all cases.

For the functional part of equation (2), we emply the same form as the functions in the FLM by Hall and Horowitz (2007). Specifically, we generate the simulation data $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ using

$$\boldsymbol{y} = (\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1} \Big(\int_0^1 \boldsymbol{x}(t)\beta(t)dt + 0.5\boldsymbol{\epsilon} \Big), \qquad \epsilon_i \sim N[0, 1],$$

where the spatial parameter ρ is assigned values of 0, 0.5, and 0.8, respectively. The

functional predictor $\boldsymbol{x}(t) = (x_1(t), x_2(t), \cdots, x_n(t))'$ is produced with values of $x_i(t)$ that are independently generated as

$$x_i(t) = \sum_{j=1}^{50} a_j Z_j \varphi_j(t),$$

where $a_j = (-1)^{j+1} j^{-\gamma/2}$ with $\gamma = 1.1$ and 2, respectively; $Z_j \sim U[-\sqrt{3}, \sqrt{3}]$ and $\varphi_j(t) = \sqrt{2}\cos(j\pi t)$. Similarly, the coefficient function $\beta(t)$ is generated according to

$$\beta(t) = \sum_{j=1}^{50} b_j \varphi_j(t),$$

where $b_1 = 0.3$ and $b_j = 4(-1)^{j+1} j^{-2}, j \geq 2$.

Note that the eigenvalues of the covariance function $\hat{K}(u, v)$ play a vital role in determining the estimation accuracy of $\beta(t)$. We consider two cases. In case 1, $\gamma = 1.1$; thus the eigenvalues are well spaced and the slope function can be accurately estimated. In case 2, $\gamma = 2$, and the closely spaced eigenvalues can cause the estimator $\hat{\beta}(t)$ to display poor performance.

The experiment is repeated 500 times in each setting. We assess the performance of the estimator $\hat{\rho}$ in terms of the mean bias and standard derivation; on the other hand, we evaluate the performance of the estimator of $\beta(t)$ in terms of the mean square error (MSE) evaluated at 100 equispaced points on $[0, 1]$, $\{t_i\}_{i=1}^{100}$; i.e.,

$$\text{MSE} = \frac{1}{100} \sum_{i=1}^{100} \left(\hat{\beta}(t_i) - \beta(t_i)\right)^2,$$

where $\hat{\beta}(t_i)$ is the estimator of $\beta(t_i)$ evaluated at t_i, which is obtained via the SFLM or the FLM. We summarize these results in Table 1.

Examination of Table 1 leads to the following conclusions.

1) When $\rho = 0$, $\widehat{\rho}$ is very small (almost zero), the estimators of $\beta(t)$ based on the proposed method and the FPCA-based method for FLM perform equally

Table 1: The empirical average biases and standard deviations (in brackets) of the estimator of ρ, denoted bias(sd), and the empirical average MSE and its standard deviations (in brackets) of $\beta(t)$ obtained via SFLM and FLM, denoted MSE_1(sd) and MSE_2(sd), respectively.

		$\gamma = 1.1$			$\gamma = 2$		
ρ	n	bias(sd)	MSE_1(sd)	MSE_2(sd)	bias(sd)	MSE_1(sd)	MSE_2(sd)
0	300	−0.0051 (0.0495)	0.0203 (0.0072)	0.0203 (0.0072)	−0.0086 (0.0628)	0.1171 (0.0239)	0.1171 (0.0239)
	500	−0.0003 (0.0428)	0.0087 (0.0030)	0.0087 (0.0029)	−0.0020 (0.0459)	0.0691 (0.0109)	0.0691 (0.0109)
	900	−0.0024 (0.0294)	0.0034 (0.0011)	0.0034 (0.0011)	0.0006 (0.0369)	0.0378 (0.0044)	0.0378 (0.0044)
0.5	300	−0.0062 (0.0457)	0.0201 (0.0071)	0.0267 (0.0101)	−0.0068 (0.0524)	0.1173 (0.0226)	0.1198 (0.0227)
	500	−0.0016 (0.0343)	0.0085 (0.0027)	0.0114 (0.0036)	−0.0037 (0.0400)	0.0689 (0.0101)	0.0702 (0.0102)
	900	−0.0034 (0.0241)	0.0034 (0.0010)	0.0047 (0.0013)	−0.0046 (0.0292)	0.0380 (0.0045)	0.0386 (0.0046)
0.8	300	−0.0062 (0.0261)	0.0202 (0.0069)	0.0836 (0.0361)	−0.0094 (0.0330)	0.1173 (0.0228)	0.1504 (0.0332)
	500	−0.0027 (0.0202)	0.0087 (0.0028)	0.0405 (0.0152)	−0.0057 (0.0245)	0.0689 (0.0110)	0.0876 (0.0161)
	900	−0.0024 (0.0149)	0.0033 (0.0011)	0.0190 (0.0058)	−0.0040 (0.0189)	0.0382 (0.0044)	0.0479 (0.0060)

well. This result is consistent with our expectations, as the SFLM reduces to the classical FLM when $\rho = 0$.

2) When $\rho \neq 0$, our proposed method produces better results than the FPCA-based method, also consistent with our expectations. The MSE of the SFLM is always smaller than the MSE of the FLM when the other settings are identical; moreover, as ρ increases, the difference in the MSE between the SFLM and the FLM also increases.

3) Regardless of the value of ρ, the MSE of $\beta(t)$ obtained using the SFLM decreases as the sample size increases. The standard deviation of $\hat{\rho}$ also displays a decreasing pattern. Moreover, the bias of $\hat{\rho}$ is small in all cases. Similar to the numerical results presented in Lee (2004), the bias of ρ is negative at all settings.

4) As the case in Hall and Horowitz (2007), $\beta(t)$ is more accurately estimated

given $\gamma = 1.1$ than $\gamma = 2$ when the other simulation parameters are held equal. The performance of the estimator $\hat{\rho}$ is also influenced by α. In the case in which $\gamma = 1.1$, the standard deviation of $\hat{\rho}$ is smaller than that of $\gamma = 2$.

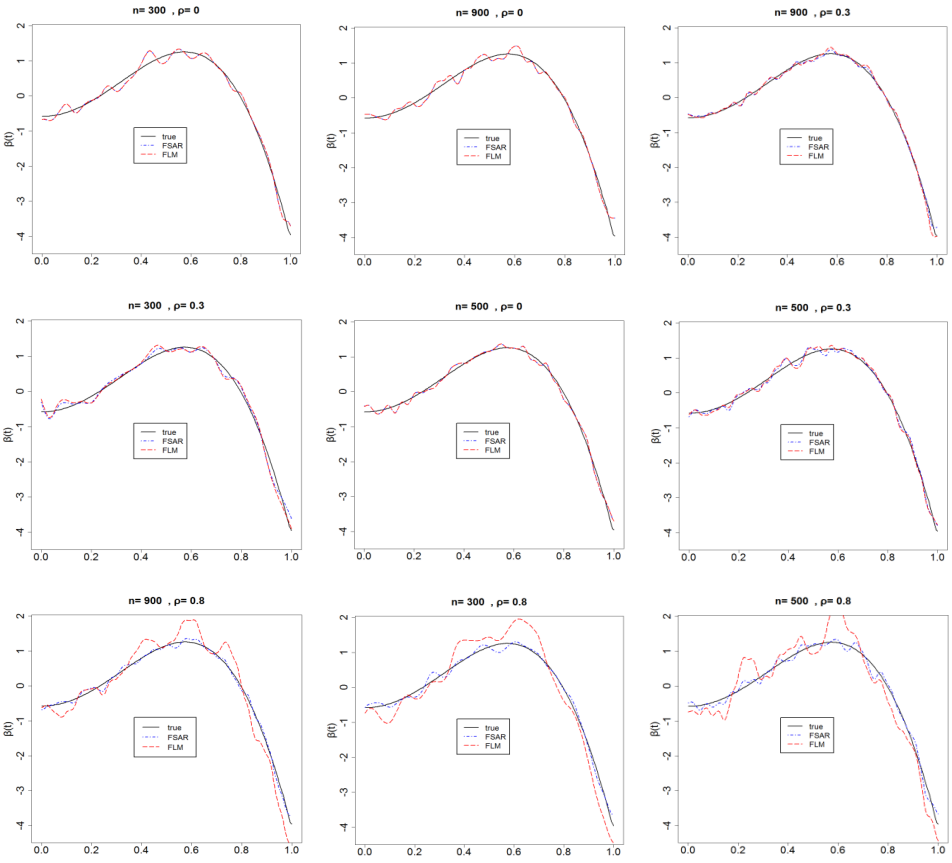

Figure 2: Estimators of $\beta(t)$ vs the true $\beta(t)$ when the sample size $n = 300, 500, 900$ and $\rho = 0, 0.5, 0.8$, respectively.

Moreover, to illustrate the performance of the estimator of $\beta(t)$ intuitively, we randomly select one result from 500 repetitions, and display the estimator $\hat{\beta}(t)$ vs. the true value of $\beta(t)$ in Figure 2 for both the SFLM and the FLM. Here, the sample size $n = 300, 500, 900$ and $\rho = 0, 0.5, 0.8$, respectively, with $\gamma = 1.1$. From Figure 2, we can obtain similar conclusions for $\beta(t)$ as those derived from Table 1.

In summary, our proposed model and estimators perform as well as the classical FLM when no spatial autocorrelation is present; on the other hand, our proposed methods

outperforms the classical methods when spatial autocorrelation is present, and the difference increases with the degree of spatial autocorrelation becoming stronger. Given these results, our proposed model and estimation procedure provide a competitive alternative for the existing methods in FDA.

5 REAL DATA ANALYSIS

In this section, we revisit the weather data presented in Section 1 to assess the application of the SFLM. Specifically, we add a record that corresponds to the weather data for 2008 derived from the China Meteorological Yearbook. Let the response y_i and the predictor $x_i(t)$ be the logarithm of the mean annual precipitation and the mean monthly temperature curve for the ith city between 2005 and 2007. We build the SFLM model as

$$y_i = \rho \sum_{i \neq i'} w_{ii'} y_{i'} + \int_0^1 x_i(t)\beta(t)dt + \epsilon_i, \tag{5.1}$$

where $w_{ii'}$ is the weight between city i and i'. We also build the FLM as

$$y_i = \int_0^1 x_i(t)\beta(t)dt + \epsilon_i \tag{5.2}$$

to enable comparison with the SFLM.

During preprocessing, we smooth the mean monthly temperatures over 3 years using the Epanechnikov Kernel. Note that the spatial correlation between temperature curves is beyond the scope of our article. Thus, we simply presume that these functional data are independent. Moreover, the spatial weight matrix is formed using the nearest k neighbours; each neighbour's weight is equal to the reciprocal of the Euclidean distance $d(i, i')$ between cities i and i'; i.e. $w_{ii'} = d(i, i')^{-1}$. Because k can take many different values, we set $k = \{2, 3, 4, 5, 6, 7, 8, 9\}$ to enable the selection of the most appropriate number. Moreover, we assume that, if the distance between two cities exceeds 15, then

these two cities are not neighbours of each other; i.e. $w_{ii'} = 0$. Figure 3 presents the locations of 34 major cities in China. Because Urumchi is located far from the other cities, we remove its record from the weather data. The matrix is row-normalized after construction.

Figure 3: The locations of 34 major cities in China

Figure 4 (top-left) shows the eigenvalues of the sample covariance function. The eigenvalues clearly decay quickly, and the first two eigenvalues account for 99% of the total variance. Therefore, we consider two candidate truncation parameters, 1 and 2. In the first case, only one covariate is contained in the truncated SFLM; in the second case, two covariates are included. Combining the 8 possible weight matrices mentioned above with these two parameters, we have 16 candidates for the truncated SFLM. Because 16 is relatively small, we compute their posterior model probabilities directly, instead of using a strategic stochastic Markov chain process to search for an appropriate model. The log-likelihood function values and Bayesian model probability ratios of each model are displayed in Table 2.

The results presented in Table 2 can be summarized as follows. The matrix with 5 nearest neighbours is the most appropriate for our model, and when $m = 2$, the SFLM yields a greater posterior probability than when $m = 1$. We display the fitting results of the SFLM and the FLM with $m = 2, k = 5$ in Table 3 and Figure 4. To assess the predictive ability of the SFLM, we apply the fitting result to the temperature

Table 2: The values of the log-likelihood function and the Bayesian model probability ratios of the 16 candidates of the truncated SFLM, denoted L and B, respectively. Specifically, the Bayesian model probability ratio is obtained by dividing the posterior probability of the current model by the posterior probability of the model for which m and k are 1 and 2.

m/k		$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$m=1$	L	-9.38	-8.54	-5.42	-3.72	-4.13	-4.10	-4.47	-4.94
	B	1	2.46	44.37	198.31	141.03	145.42	108.51	69.89
$m=2$	L	-4.42	-3.28	-1.51	-1.11	-1.90	-2.15	-2.25	-2.50
	B	90.39	302.32	1755.32	2253.14	1163.4	943.31	887.44	695.95

Table 3: The fitting and prediction results of the SFLM and the FLM. Here, the fitting error and prediction error are evaluated separately in terms of the mean square error of the fitted values and the prediction values.

models	$\hat{\rho}$	Moran's I statistic(residuals)	MSE(fitted error)	MSE(prediction error)
FLM	—	0.41	0.33	0.12
SFLM	0.58	0.15	0.24	0.10

observations made in 2008 to determine the annual precipitation in the same year. The fitting results and prediction errors are displayed in Table 3 and Figure 4.

Recall from Section 1 that Figure 1 reflects significant spatial autocorrelation among the annual precipitation values for the different cities, and the FLM can not efficiently address this correlation. The Moran's I scatterplot of residuals of the SFLM in Figure 4 suggests that a majority of the spatial autocorrelation in responses has been removed. We can also conclude from Figure 4 that the SFLM has reduced the fitting error substantially when compared to the FLM.

As for the estimated parameters, $\hat{\rho}$ is 0.58, and the corresponding P value is smaller than 0.001; the slope functions of the two models are provided in Figure 4 (to the right). The two curves of the estimated $\beta(t)$ have similar shapes, and the $\hat{\beta}(t)$ of the SFLM is smoother than that of the FLM. We conclude that precipitation is much more strongly

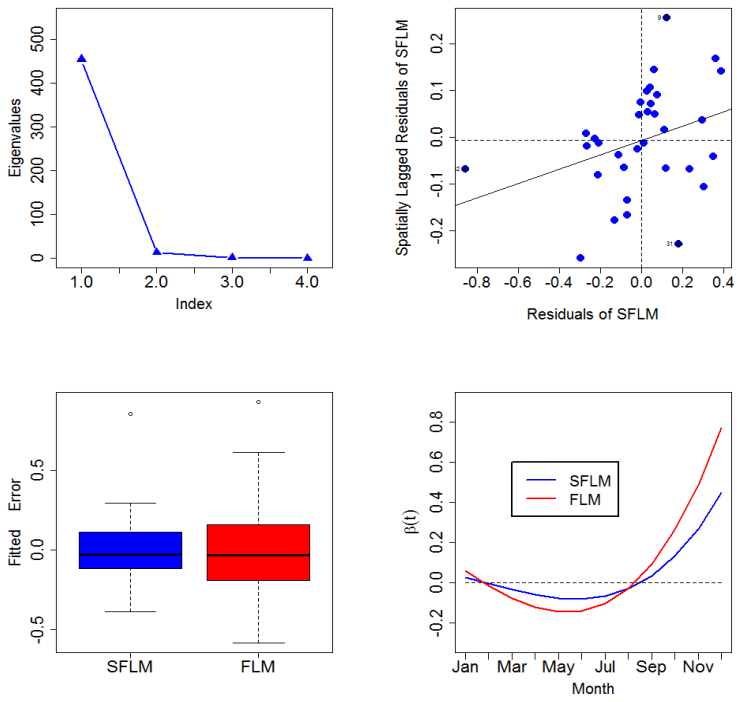

Figure 4: Eigenvalues of sample covariance function (top left), Moran I scatter plot of residuals of SFLM (top right), fitted error of SFLM and FLM (bottom left), and estimated $\beta(t)$ of SFLM and FLM (bottom right), respectively.

influenced by temperature during the winter than in the other seasons; moreover, under SFLM, the precipitation received by each city is less affected by temperature over the year as a whole.

To illustrate the superiority of the SFLM relative to the FLM, we also compare their predictive abilities. We use the data from 2005 to 2007 to predict the precipitation in 2008. The MSE values of the predicted y_i under the SFLM and FLM are 0.10 and 0.12, respectively, which reflects substantial improvement.

6 CONCLUSION AND DISCUSSION

The FLM is popular in studies of links between a scalar response and functional predictors. However, the existing FLM cannot address the dependencies that arise due to

the presence of a network structure. We propose a powerful spatial functional linear model that integrates the advantages of FLM in handling high-dimensional data and SAR model in coping with spatial dependencies. A simple estimation method is developed to obtain the estimators of the spatial autoregressive parameter and the slope function. Our simulation study demonstrates the consistency of the proposed estimators. In particular, the new estimators perform as well as the FPCA-based estimator of the FLM when the spatial autoregressive parameter equals zero; on the other hand, the new methods outperform the existing ones when spatial autocorrelation is present. An examination of a real dataset demonstrates the superiority of the SFLM over the FLM. From this perspective, our proposed model and estimation procedure represent competitive alternatives to the FLM.

Note that network structure is indeed what is considered in the SFLM, and the concept of network-structured data is more general than the spatially correlated data considered in this article. For example, social network data display network structure, but they are not spatially correlated. In fact, the weather data studied here are spatially dependent in terms of their network structure, which provides a new perspective on spatial functional data. Moreover, we discuss the SFLM under the assumption that only one functional predictor is involved. However, the SFLM with multiple functional predictors also deserve attention. Based on the estimation method introduced in Section 3, the methods presented here can easily be generalized to the SFLM with multiple functional predictors. Functional variable selection can then be conducted. These steps are beyond the scope of the current paper and will be investigated further in subsequent work.

Acknowledgements

This research was financially supported by the National Natural Science Foundation of China under grant nos. 71420107025 and 11701023.

References

Aguilera-Morillo, M. C., Durbn, M., and Aguilera, A. M. (2016). Prediction of functional data with spatial dependence: a penalized approach. *Stochastic Environmental Research & Risk Assessment*, pages 1–16.

Aneiros-Pérez, G. and Vieu, P. (2006). Semi-functional partial linear regression. *Statistics and Probability Letters*, **76**(11), 1102–1110.

Brockhaus, S., Scheipl, F., Hothorn, T., and Greven, S. (2015). The functional linear array model. *Statistical Modelling*, **15**(3), 279–300.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, **34**(5), 2159–2179.

Case, A. C. (1991). Spatial patterns in household demand. *Econometrica*, **59**(4), 953–965.

Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2004). Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, **119**(1), 1–6.

Ferraty, F., Goia, A., Salinelli, E., and Vieu, P. (2013). Functional projection pursuit regression. *test*, **22**(2), 293–320.

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

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer Science & Business Media.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 411–432.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, **40**(2), 509C533.

Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, **72**(6), 1899–1925.

Lee, L. F. (2007). Gmm and 2sls estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, **137**(2), 489–514.

Lesage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. CRC Press.

Lesage, J. P. and Parent, O. (2007). Bayesian model averaging for spatial econometric models. *Geographical Analysis*, **39**(3), 241C267.

Liu, B., Wang, L., and Cao, J. (2017). Estimating functional linear mixed-effects regression models. *Computational Statistics and Data Analysis*, **106**, 153–164.

Olubusoye, O. E., Korter, G. O., and Salisu, A. A. (2016). Modelling road traffic crashes using spatial autoregressive model with additional endogenous variable. *Statistics in Transition New*, **17**, 659–670.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, **70**(349), 120–126.

Paganoni, A. M. and Sangalli, L. M. (2017). Functional regression models: Some directions of future research. *Statistical Modelling*, **17**(1-2), 94–99.

Pinedaros, W. and Giraldo, R. (2016). Functional sar model.

Qu, X. and Lee, L. F. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics*, **184**(2), 209–232.

Ramn, G., Pedro, D., and Jorge, M. (2017). Spatial prediction of a scalar variable based on data of a functional random field. *Comunicaciones en Estadstica*, **10**(2), 315–344.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. Springer.

Topa, G. (2001). Social interactions, local spillovers and unemployment. *Review of Economic Studies*, **68**(2), 261–295.

Zhou, J., Wang, N. Y., and Wang, N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, **23**(1), 25–50.