Spatial Functional Linear Model and Its Estimation Method

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

[1]School of Economics and Management,Beihang University,Beijing,China
[2]CEDRIC, CNAM, Paris
[3]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing [4]Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations,Beijing,China

Dec., 2019

Contents

1.1 functional data

What is functional data?

Functional data is a type of data recorded at a high frequency , and almost can be regarded as continuously observed.

trajectory data, weather data and stock data.

functional data are inherently

- infinite-dimensional
- rich in information, derivative
- biology, medical sciences, econometrics

1.2 functional linear model

Functional linear model (FLM) becomes more and more popular owing to its adaptability to problems which are hard to deal with in the framework of scalar and vector observations.

Cai and Hall (2006); Hall and Horowitz (2007); Ferraty et al. (2013)

$$Y = \alpha + \int_{\Gamma} X(t)\beta(t)dt + \epsilon, \tag{1}$$

- Y : a scalar response variable
- $X(t)$: a second-order stochastic process on a compact interval Γ, $E(X(t)) = 0$, $E(\int_{\Gamma} X^2(t)dt) < \infty$
- $\beta(t)$: unknown slope function
- ϵ : random error term independent of $X(t)$ with zero mean and finite variance

1.2 functional linear model

Extensions of FLM have been studied to address specific problems.

- James (2002)put forward functional logistic regression and functional censored regression
- Aneiros-Prez and Vieu (2006) constructed a semi-functional partial linear model
- Ferraty et al. (2013) generalized the FLM to functional projection pursuit regression that allows for more interpretability
- Liu et al. (2017) presented a functional linear mixed model

1.3 functional linear model

- *Firstly*, all aforementioned methods assume that the individuals are mutually independent.

  Secondly, relation information among individuals with advance of information technology can be easily collected.

  And as one typical relational information, the network-structured data are becoming increasing available.

1.3 functional linear model

- *Firstly*, all aforementioned methods assume that the individuals are mutually independent.

  Secondly, relation information among individuals with advance of information technology can be easily collected.

  And as one typical relational information, the network-structured data are becoming increasing available.

- If we use FLM or the aforementioned variations of FLM to model this kind of data

1.3 functional linear model

- *Firstly*, all aforementioned methods assume that the individuals are mutually independent.

  Secondly, relation information among individuals with advance of information technology can be easily collected.

  And as one typical relational information, the network-structured data are becoming increasing available.

- If we use FLM or the aforementioned variations of FLM to model this kind of data

  1. the information contained in the data may not be fully exploited

1.3 functional linear model

- *Firstly*, all aforementioned methods assume that the individuals are mutually independent.

  Secondly, relation information among individuals with advance of information technology can be easily collected.

  And as one typical relational information, the network-structured data are becoming increasing available.

- If we use FLM or the aforementioned variations of FLM to model this kind of data

  1. the information contained in the data may not be fully exploited
  2. The inference may be misleading when ignoring such network effects

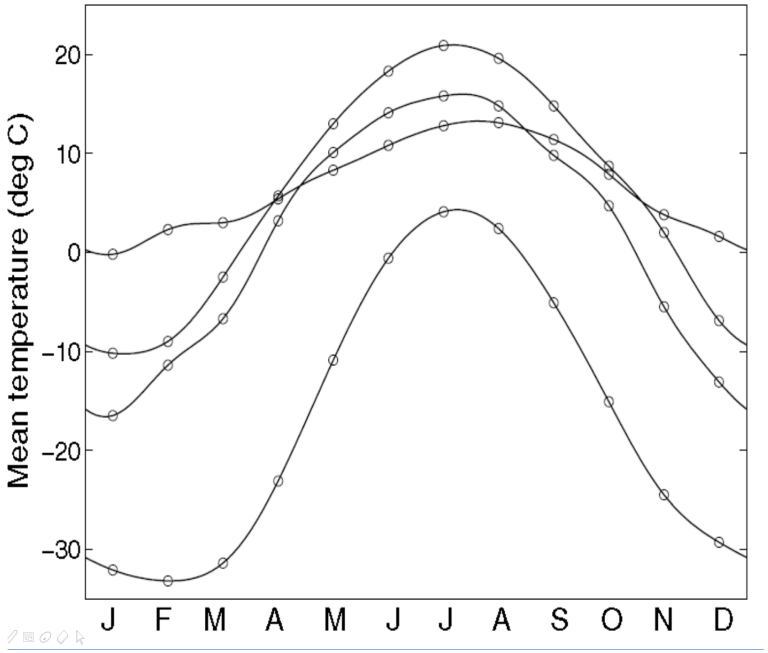
1.4 motivating example

Weather data of temperature and precipitation in cities of China

- Data source : China Meteorological Yearbook
- Description : monthly temperature and precipitation in 34 major cities of China between 2005 and 2007
- Problem : effect of temperature on the precipitation over three years

Y is the logarithm of *mean annual precipitation*
$X(t)$ is the *mean monthly temperature*

1.4 motivating example



monthly temperature in a year

key cities

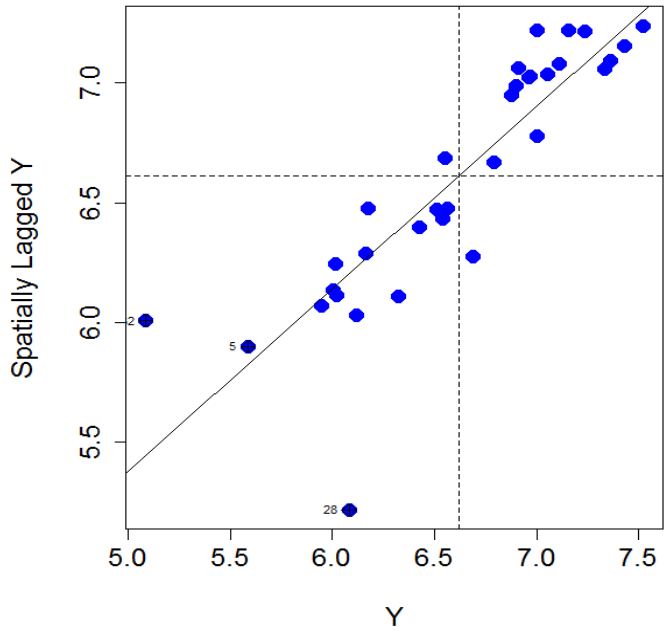
1.4 motivating example

We adopt Moran I statistic to test whether there exists spatial autocorrelation between the responses or not

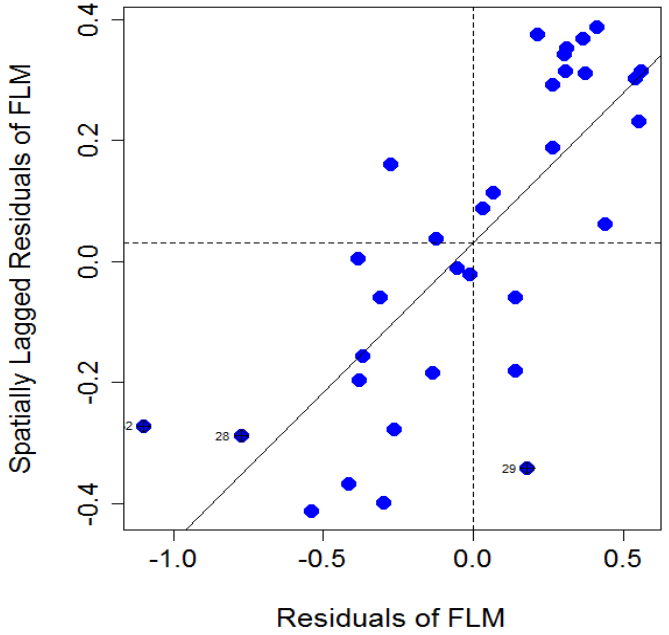
the original precipitation	the residuals of FLM
Moran I statistic $=$ **0.7**	Moran I statistic $=$ **0.5**
P value $<$ 0.000 001	P value $<$ 0.000 001

we can observe that the responses as well as the residuals of FLM showed a significance spatial autocorrelation, which means that FLM may be no longer appropriate for such data.

1.4 motivating example



moran I plot of precipitation



Moran I plot of residuals of FLM

1.4 motivating example

When the predictor is scalar, the spatial autoregressive model (SAR) is the popularly used model to accommodate the dependence brought by the network structure.

$$\mathbf{y} = \alpha\boldsymbol{\tau}_n + \rho\mathbf{W}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

- a spatial weight matrix is adopted to denote adjacent relations
- an unknown spatial autoregressive parameter is used to reflect the strength of neighboring effects

2. spatial functional linear model

Borrowing thoughts from SAR model, we propose a new model which is Spatial Functional Linear Model (SFLM).

Model Assumption

- Following Qu and Lee (2015), we consider the spatial process located on a 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.
- $x_i(t)$'s are 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 $E(\int_\Gamma X^2(t)dt) < \infty$.

$$\mathbf{y} = \alpha\boldsymbol{\tau}_n + \rho\mathbf{W}\mathbf{y} + \int_0^1 \mathbf{x}(t)\beta(t)dt + \boldsymbol{\epsilon} \qquad (3)$$

2. spatial functional linear model

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

- $\alpha\boldsymbol{\tau}_n$ is the intercept, α denotes a scalar parameter
- ρ is the unknown spatial autocorrelation parameter takes value from $[0, 1)$
- $\mathbf{W} = (w_{ii'})_{n\times n}$ is a pre-specified spatial weight matrix
- $\beta(t)$ is a square integrable coefficient function defined on $[0, 1]$
- ϵ is the noise term independent of $\mathbf{x(t)}$, follows a multivariate normal distribution with zero mean and a constant diagonal covariance matrix $\sigma^2\mathbf{I}_n$

2. spatial functional linear model

notes of spatial functional linear model

- the spatial weight matrix $\mathbf{W}$ is exogenous
- ρ is a scale parameter measuring the power of the neighbor impact
- the SFLM can degenerate into the following two models
  - functional linear model (FLM) when $\rho = 0$
  - spatial autoregressive model (SAR) when $\mathbf{x(t)}$ is free of t

2. spatial functional linear model

we can reformulate model (3) as the following equivalent expression as well,

$$\mathbf{y} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}(\alpha\boldsymbol{\tau}_n + \int_0^1 \mathbf{x}(\mathbf{t})\beta(t)dt + \boldsymbol{\epsilon}) \qquad (4)$$

- y_i is influenced by its neighbors' covariates $x_{i'}(t)$s, $i' \neq i$
- Gauss-Markov assumption is violated, the least square estimator (OLS) is not adequate

3. estimation method

1. Basis Expansion based on FPCA
2. MLE for truncated SFLM

3. estimation method

Basis Expansion based on FPCA

$$K(s,t) = \mathrm{Cov}(X(t), X(s)).$$
$$K(s,t) = \sum_{j=1}^{\infty} k_j \varphi_j(s)\varphi_j(t) \quad , \quad k_1 > k_2 > \cdots > 0.$$
$$X(t) = \sum_{j=1}^{\infty} a_j \varphi_j(t).$$

For the observation $\{y_i, x_i(t)\}_{i=1}^n$,

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

Basis Expansion based on FPCA

Also we can obtain that

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

$$\hat{a}_{ij} = \int_0^1 x_i(t) \hat{\varphi}_j(t) \ , \ x_i(t) = \sum_{j=1}^{\infty} \hat{a}_{ij} \hat{\varphi}_j(t)$$

$$b_j = \int_0^1 \beta(t) \hat{\varphi}_j(t) dt \ , \ \beta(t) = \sum_{j=1}^{\infty} b_j \hat{\varphi}_j(t)$$

.

$$\mathbf{y} = \alpha \boldsymbol{\tau}_n + \rho \mathbf{W} \mathbf{y} + \sum_{j=1}^{\infty} \hat{\mathbf{a}}_\mathbf{j} b_j + \boldsymbol{\epsilon}, \tag{5}$$

Basis Expansion based on FPCA

Using the criterion, percentage of variance explained (PVE) for predictors to choose number of PCs.
The truncated model is

$$\mathbf{y} \approx \alpha \boldsymbol{\tau}_n + \rho \mathbf{W}\mathbf{y} + \sum_{j=1}^{m} \hat{\mathbf{a}}_\mathbf{j} b_j + \boldsymbol{\epsilon}. \tag{6}$$

Based on estimated $\hat{b}_j$, we can reconstruct $\hat{\beta}(t)$ by

$$\hat{\beta}(t) = \sum_{j=1}^{m} \hat{b}_j \hat{\varphi}_j(t) \tag{7}$$

MLE for truncated SFLM

Defining $\mathbf{A} = (\hat{a}_{ij})_{n\times m}$, $\mathbf{b} = (b_1, b_2, \cdots, b_m)'$, $\mathbf{Z} = (\boldsymbol{\tau}_n, \mathbf{A})$ and $\boldsymbol{\delta} = (\alpha, \mathbf{b}')'$, we can write the truncated model (6) as

$$\mathbf{y} \approx \rho\mathbf{W}\mathbf{y} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}. \tag{8}$$

Since model (8) is similar to the popular SAR model, we adopt the maximum likelihood estimation method to get parameters.

3. estimation method

1. Represent the functional predictor and slope function by functional principle component basis, then the estimation procedure is simplified as the remaining process is similar to the estimation problem of a SAR model.
2. Get estimators of unknown parameters in the truncated SFLM obtained from Step 1.
3. Get the estimator of $\beta(t)$ in SFLM. The slope function is constructed with the FPC basis in Step 1 and the estimated coefficients in Step 2. Other estimators are directly obtained from Step 2.

4. simulation study

- As for spatial scenario, we adopt the rook matrix by randomly apportioning n agents on a regular square grid of cells.
- And for the functional part in Model (2), we take the same form as functions in FLM by Hall and Horowitz (2007).

$$Y = \rho WY + \int_0^1 X(t)\beta(t)dt + 0.5\epsilon$$

$$X(t) = \sum_{j=1}^{50} a_j Z_j \varphi_j(t) \qquad \beta(t) = \sum_{j=1}^{50} b_j \varphi_j(t)$$

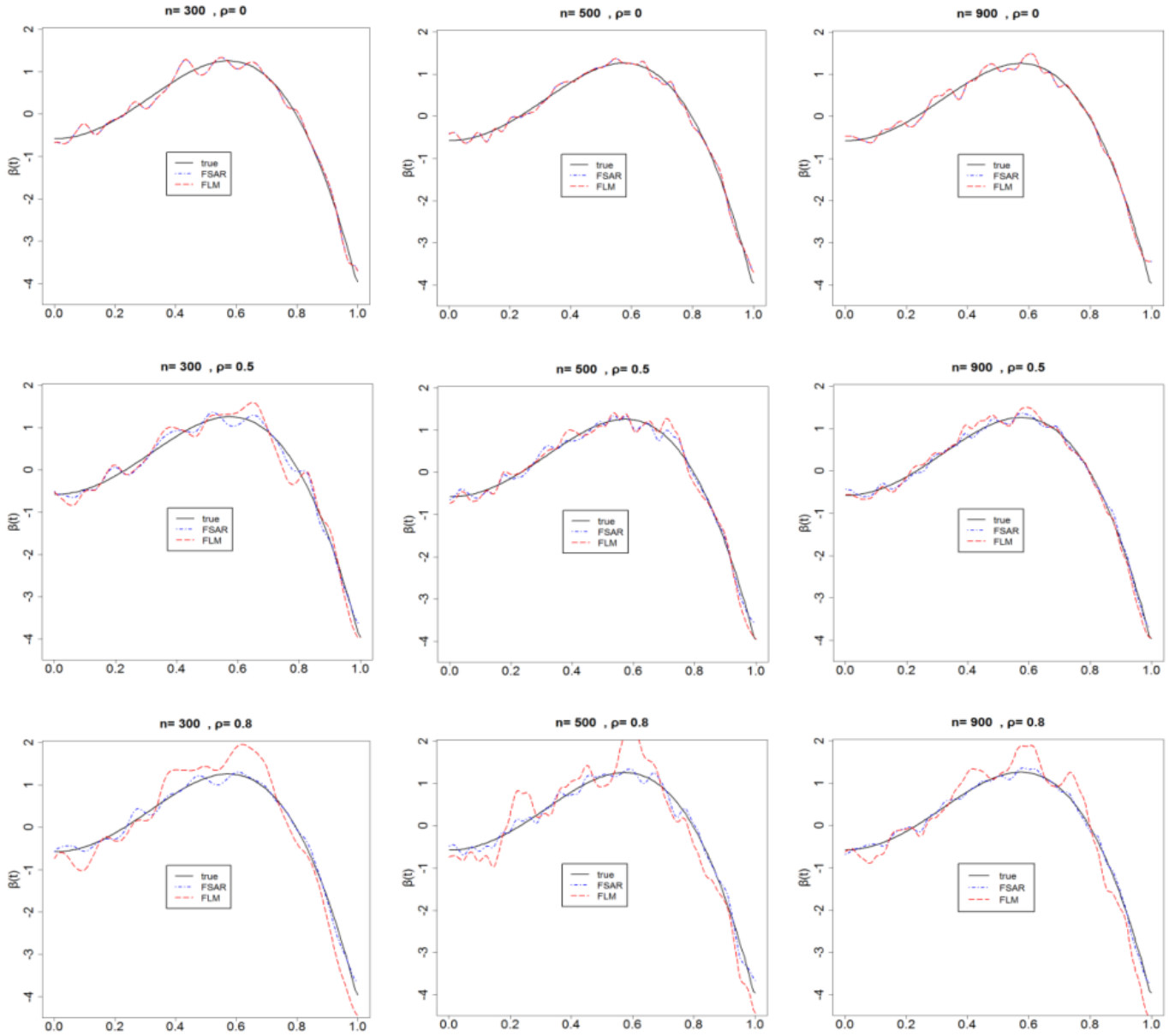
$a_j = (-1)^{j+1} j^{-\gamma/2}$ $\quad Z_j \sim U[-\sqrt{3}, \sqrt{3}]$ $\quad \epsilon \sim N[0,1]$
$b_1 = 0.3$ $\quad b_j = 4(-1)^{j+1} j^{-2}, j \geq 2$ $\quad \varphi_j(t) = \sqrt{2}cos(j\pi t)$
$\rho = \{0, 0.5, 0.8\}$ $\quad \gamma = \{1.1, 2\}$ $\quad n = 300, 500, 900$

4. simulation study

		$\gamma = 1.1$			$\gamma = 2$		
ρ	n	bias(sd)	MSE_1(sd)	MSE_2(sd)	bias(sd)	MSE_1(sd)	MSE_2(sd)
0	300	−0.0051 (0.050)	0.0203 (0.007)	0.0203 (0.007)	−0.0086 (0.063)	0.1171 (0.024)	0.1171 (0.024)
	500	−0.0003 (0.043)	0.0087 (0.003)	0.0087 (0.003)	−0.0020 (0.046)	0.0691 (0.011)	0.0691 (0.011)
	900	−0.0024 (0.029)	0.0034 (0.001)	0.0034 (0.001)	−0.0024 (0.029)	0.0034 (0.001)	0.0034 (0.001)
0.5	300	−0.0062 (0.046)	0.0201 (0.007)	0.0267 (0.010)	−0.0068 (0.052)	0.1170 (0.023)	0.1198 (0.023)
	500	−0.0016 (0.034)	0.0085 (0.003)	0.0114 (0.004)	−0.0037 (0.040)	0.0689 (0.010)	0.0702 (0.010)
	900	−0.0034 (0.024)	0.0034 (0.001)	0.0047 (0.001)	−0.0046 (0.029)	0.0380 (0.005)	0.0386 (0.005)
0.8	300	−0.0062 (0.026)	0.0202 (0.007)	0.0836 (0.036)	−0.0094 (0.033)	0.1173 (0.023)	0.1504 (0.033)
	500	−0.0027 (0.020)	0.0087 (0.003)	0.0405 (0.015)	−0.0057 (0.025)	0.0689 (0.011)	0.0876 (0.016)
	900	−0.0024 (0.015)	0.0033 (0.001)	0.0190 (0.006)	−0.0040 (0.019)	0.0382 (0.004)	0.0479 (0.006)

$$MSE_1 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\beta}_{SFLM}(t_i) - \beta(t_i))^2 \ , \ MSE_2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\beta}_{FLM}(t_i) - \beta(t_i))^2$$

4. simulation study



4. simulation study

1. When $\rho = 0$, our proposed method and FPCA based method perform equally well.
2. When $\rho \neq 0$, our proposed method behaves better than FPCA based method.
3. No matter what ρ equals to, MSE_1 of $\beta(t)$ decreases as sample size increases. And the standard deviation of $\widehat{\rho}$ has a decreasing pattern.
4. As the case in Hall and Horowitz (2007), $\beta(t)$ is better estimated with $\gamma = 1.1$ than $\gamma = 2$ when other simulation parameters equal.

5. weather data

We perform the SFLM and FLM on the average of weather data from 2005 to 2007 and *compare their efficiency*. The SFLM is formulated as

$$y_i = \rho \sum_{i \neq i'} w_{ii'} y_{i'} + \int_0^1 x_i(t)\beta(t)dt + \epsilon_i, \tag{9}$$

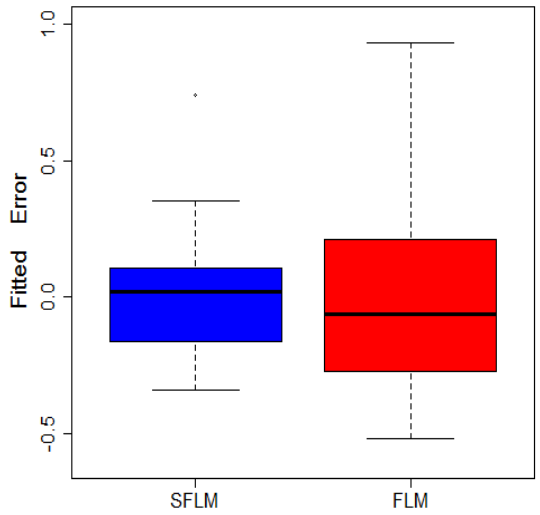
where $w_{ii'}$ is the weight between city i and i'. We also built the FLM as

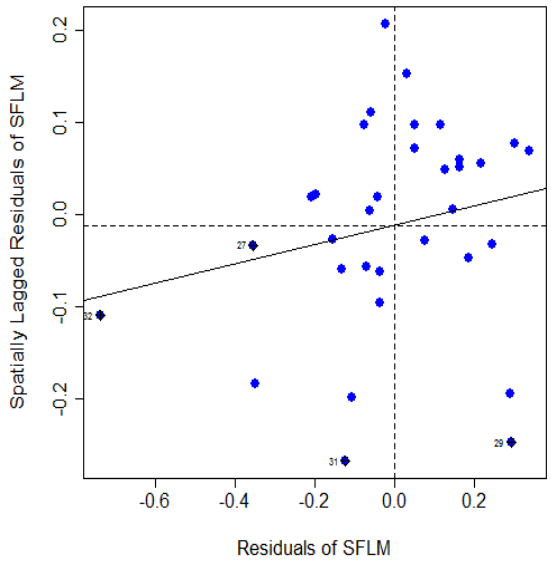
$$y_i = \int_0^1 x_i(t)\beta(t)dt + \epsilon_i \tag{10}$$

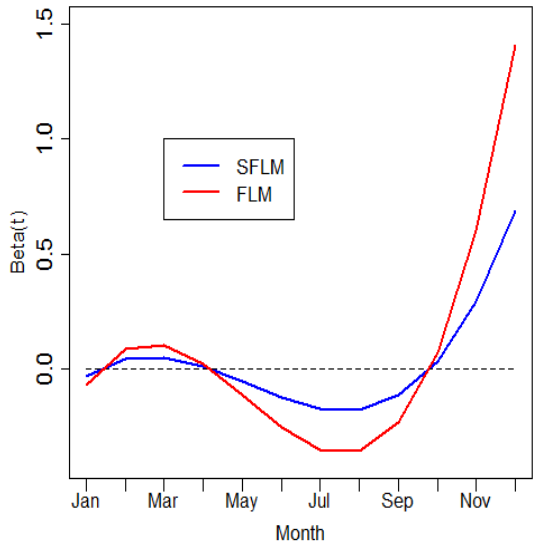
Main steps

1. we smoothed the mean monthly temperature over 3 years by Epanechnikov Kernel.
2. the spatial weight matrix was formed by nearest 5 neighbors, each neighbor's weight equalling reciprocal of Euclidean distance $d(i, i')$ between cities i and i'.
3. Urumchi's location is far from other cities, we removed its record in weather data.
4. We perform the SFLM in model (10) and FLM in model (11) to get estimators.
5. Then we apply the fitting result to temperature observations in 2008 to predict annual precipitation.

5. weather data







We conclude that the precipitation is much more influenced by temperature in winter than in other seasons. And under SFLM the precipitation of each city is less affected by temperature during the whole year.

5. weather data

Table 2. The fitting and predicting results

models	$\widehat{\rho}$	fitted error	predicted error	Moran I statistic
FLM	0.66	0.40	0.16	0.50
SFLM	—	0.25	0.13	0.13

6. conclusion

1. We propose a powerful spatial functional linear model which integrates the advantages of FLM in dealing with high dimensional data and SAR model in coping with spatial dependence.

6. conclusion

1. We propose a powerful spatial functional linear model which integrates the advantages of FLM in dealing with high dimensional data and SAR model in coping with spatial dependence.
2. A simple estimation method is developed to obtain estimators

6. conclusion

1. We propose a powerful spatial functional linear model which integrates the advantages of FLM in dealing with high dimensional data and SAR model in coping with spatial dependence.
2. A simple estimation method is developed to obtain estimators
3. Our simulation study demonstrate the consistency of the proposed estimators.

6. conclusion

1. We propose a powerful spatial functional linear model which integrates the advantages of FLM in dealing with high dimensional data and SAR model in coping with spatial dependence.
2. A simple estimation method is developed to obtain estimators
3. Our simulation study demonstrate the consistency of the proposed estimators.
4. A real dataset study demonstrates superiority of SFLM over FLM.

6. conclusion

1. We propose a powerful spatial functional linear model which integrates the advantages of FLM in dealing with high dimensional data and SAR model in coping with spatial dependence.
2. A simple estimation method is developed to obtain estimators
3. Our simulation study demonstrate the consistency of the proposed estimators.
4. A real dataset study demonstrates superiority of SFLM over FLM.
5. The SFLM with multiple functional predictors can be also obtained to deal with problem in practice.

6. conclusion

1. We propose a powerful spatial functional linear model which integrates the advantages of FLM in dealing with high dimensional data and SAR model in coping with spatial dependence.
2. A simple estimation method is developed to obtain estimators
3. Our simulation study demonstrate the consistency of the proposed estimators.
4. A real dataset study demonstrates superiority of SFLM over FLM.
5. The SFLM with multiple functional predictors can be also obtained to deal with problem in practice.
6. We can also conduct functional variable selection in further research.

Main Reference

[1] Cai, T. and Hall, P. (2006). Prediction in functional linear regression. The Annals of Statistics, 34(5), 21592179.

[2] Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. Annals of Statistics, 35(1), 7091.

[3] Qu, X. and Lee, L. F. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. Journal of Econometrics, 184(2), 209232.

Thank you very much !