DOI:10.19789/j.1004-9398.2023.06.005
文献引用:惠姣姣.半参数函数型线性空间自回归模型的统计推断及应用[J].首都师范大学学报(自然科学版),2023,44(6):20-29. HUI J J. Statistical inference and application of semi-parametric functional linear spatial autoregressive model[J]. Journal of Capital Normal University (Natural Science Edition), 2023, 44(6): 20-29.

半参数函数型线性空间自回归模型的统计推断及应用*

惠　姣　姣**

(宝鸡文理学院数学与信息科学学院,陕西 宝鸡　721013)

摘要:为了解决自变量为函数型数据、响应变量为标量数据、协变量为矩阵型数据且其分布未知的问题,本文提出半参数函数型线性空间自回归模型(SP-FLSARM)。首先利用函数型主成分分析(FPCA)处理函数型自变量,利用B样条估计非参数连接函数;其次利用极大似然估计(MLE)得到未知参数的估计值;最后针对西班牙气象台的温度数据建立SP-FLSARM模型。相较于函数型线性空间自回归模型(FLSARM)和函数型线性模型(FLM),SP-FLSARM的方差和残差方差比(V_R)最小,即SP-FLSARM的模型拟合效果最好。

关键词:SP-FLSARM;FPCA;B样条;MLE;西班牙气象台温度数据
中图分类号:O212.1　　**文献标识码**:A

Statistical inference and application of semi-parametric functional linear spatial autoregressive model*

HUI Jiaojiao**

(School of Mathematics and Information Science, Baoji University of Arts and Sciences, Baoji Shaanxi　721013)

Abstract: In order to solve the problem that the independent variable is functional data, the response variable is scalar data, the covariant is matrix data and its distribution is unknown, this paper proposes a semi-parametric functional linear spatial autoregressive model (SP-FLSARM). Firstly, the functional independent variable is processed by functional principal component analysis (FPCA), the non-parametric connection function is estimated by B-spline. Secondly, the estimated value of the unknown parameter is obtained by maximum likelihood estimation (MLE). Finally, the SP-FLSARM model is established based on the temperature data of the Spanish meteorological observatory. Compared with functional linear spatial autoregressive model(FLSARM) and functional linear model (FLM), the variance and variance ratio (V_R) of SP-FLSARM are the smallest, the fitting effect of SP-FLSARM is the best.

Keywords: SP-FLSARM; FPCA; B-splines; MLE; meteorological observatory data of Spanish
CLC: O212.1　　**DC**: A

收稿日期:2022-05-09
* 陕西省教育厅自然科学研究项目(202310592)
** 通信作者:2533741876@qq.com

0 引 言

随着空间遥感技术的快速发展和地理信息系统的日益完善,空间数据呈爆炸式增长,其收集方式也呈现多样化,这使研究人员可以更加快捷地获取空间数据,但接下来如何分析空间数据一直是业界关注的焦点。Tobler[1]认为具有空间效应的数据不满足样本独立性的假设,如果对这样的数据仍使用传统的计量方法进行统计分析,那么必然会造成较大的误差,因此,空间计量模型受到了学者们的广泛关注。与此同时,日常中容易收集到的天气数据、股票数据等函数型数据的统计分析也被广泛应用到生物学、医学、气象学和经济学等领域,通常用函数型线性模型(functional linear model,FLM)[2]处理函数型数据。

数据形式逐渐多样化,当函数型数据中融入空间效应可得到空间函数型数据。研究人员为了处理这类数据,提出了几种函数型线性空间计量模型,如函数型线性空间自回归模型(functional linear spatial autoregressive model,FLSARM)、函数型线性空间误差模型(functional linear spatial error model,FLSEM)、基于空间效应的函数型条件自回归模型(conditional autoregressive model,CARM)等,这几种模型都指定了一个参数回归函数,有固定的模型表达式。但在一些特殊的情况下,有些数据的分布是无从得知的,需要借助非参数回归(non-parametric regression,NPR)来进行估计。NPR的形式较为自由、对总体分布的限制条件较为宽泛,具有稳健性高、精度高、适应性强的特点,能够很好地解决非齐次、非线性问题[3]。在普通的空间计量模型中融入NPR,便可得到半参数空间计量模型。

目前,FLM、半参数空间自回归模型(semi-parametric spatial autoregressive model,SP-SARM)和函数型线性空间计量模型的理论工作已发展的相对成熟,研究人员在不断地提高模型的估计精度。然而,当自变量为函数型数据、响应变量为标量、协变量为矩阵型数据且其分布未知时,遇到这样的实际问题又该如何建立模型并估计参数呢?目前仍未有文献提及相应的模型处理方法。鉴于此,本文考虑将SP-SARM与FLM相结合,最终构建出半参数函数型线性空间自回归模型(semi-parametric functional linear spatial autoregressive model,SP-FLSARM)。

为了有效地处理空间函数型数据,Aw和Cabral[4]将Anselin[5]提出的空间自回归模型(spatial autoregressive model,SARM)与经典的FLM[6]结合,构造出函数型空间滞后模型(functional spatial lag model,FSLM),在截断FSLM的基础上,首先利用函数型主成分分析(functional principal component analysis,FPCA)处理函数型数据,再利用Bayesian MCMC技术估计模型的未知参数,从而得到该模型的Bayesian估计;Zhang等[7]建立了一个基于空间效应的CARM,仿真研究表明在模型中考虑空间相关性可以提高FLM的性能,利用该模型对全器官组织学和遗传图谱数据进行实证分析,结果表明该模型在处理空间函数型数据时效果良好;Pinda-Ríos等[8]构建了FLSARM,首先分别利用FPCA和B样条处理函数型数据,其次利用最小二乘法和极大似然法估计参数,最后利用该模型对哥伦比亚中心地区117个镇的空间函数型数据进行分析,结果显示该模型能够较为精确地估计出空间自相关系数和斜率函数;Aw和Cabral[9]构建了空间自回归组合模型(spatial autoregressive combined model,SACM),首先利用FPCA处理函数型数据,并利用极大似然估计(maximum likelihood estimation,MLE)估计SACM的模型参数,该模型可广泛应用于地理学、经济学、人口学和区域科学等领域;胡锡健[10]提出部分函数型空间自回归(partial functional spatial auto-regressive,PFSAR)模型,并通过函数型主成分基函数法对模型中函数型部分截断处理,再通过MLE对展开后的模型进行参数估计。

以上研究都有固定的模型表达式,但在实际中,有些变量的分布无从得知,此时就可以考虑借助NPR来进行估计。Baíllo和Grané[11]将NPR与FLM相结合,利用FPCA处理了函数型数据,并用局部线性回归估计未知参数;Wang等[12]将NPR与SARM结合,提出了SP-SARM,并利用局部线性回归估计了未知参数;Su和Jin[13]利用拟极大似然估计(profiled quasi-maximum likelihood estimator,PQMLE)对SP-SARM中的未知参数进行估计,并给出了依赖于权重矩阵的有限维参数的收敛速度。这些都为后续的研究奠定了理论基础。

本文基于SP-SARM与FLM,针对自变量为函数型数据、响应变量为标量、协变量为矩阵型数据的情

形,构建了SP-FLSARM。首先,利用MLE对未知参数进行估计;其次,对SP-FLSARM和其参数估计结果进行500次模拟试验,以此评判SP-FLSARM的参数估计效果;最后,利用SP-FLSARM、FLSARM和FLM对西班牙气象台的光照数据建模分析。

1 半参数函数型线性空间自回归模型

1.1 预备知识

首先介绍本文建立模型的基础,SARM、FLM、FLSARM和SP-SARM。

SARM的一般形式[14]为

$$\begin{cases}\boldsymbol{Y}=\alpha\boldsymbol{\tau}_n+\rho\boldsymbol{W}\boldsymbol{Y}+\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon}\overset{i.i.d}{\sim}N(0,\sigma^2\boldsymbol{I}_n),\end{cases}\tag{1.1}$$

式中:$\boldsymbol{Y}$是$n\times1$维响应变量;α是截距;$\boldsymbol{\tau}_n$是所有元素均为1的n维向量;ρ是空间自相关系数;$\boldsymbol{W}$是行标准化后的空间权重矩阵;$\boldsymbol{X}$是$n\times k$维自变量,$\boldsymbol{\beta}$是$k\times1$维回归系数向量;$\boldsymbol{\varepsilon}$是随机误差项。

FLM的一般形式[15]为

$$\begin{cases}\boldsymbol{Y}=\alpha\boldsymbol{\tau}_n+\int_{\Gamma}\boldsymbol{X}(t)\beta(t)\mathrm{d}t+\boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon}\overset{iid}{\sim}N(0,\sigma^2\boldsymbol{I}_n),\end{cases}\tag{1.2}$$

式中:$\boldsymbol{Y}$是$n\times1$维响应变量;α是截距,$\boldsymbol{\tau}_n$是所有元素均为1的n维向量;$\boldsymbol{X}(t)$是紧集Γ上的二阶随机过程且满足$E(\boldsymbol{X}(t))=0$、$E(\int_{\Gamma}\boldsymbol{X}^2(t)\mathrm{d}t)<\infty$;$\beta(t)$是未知的斜率函数;$\boldsymbol{\varepsilon}$是随机误差且与$\boldsymbol{X}(t)$相互独立。

FLSARM的一般形式为

$$\begin{cases}\boldsymbol{Y}=\rho\boldsymbol{W}\boldsymbol{Y}+\int_{\Gamma}\boldsymbol{X}(t)\beta(t)\mathrm{d}t+\boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon}\overset{iid}{\sim}N(0,\sigma^2\boldsymbol{I}_n),\end{cases}\tag{1.3}$$

式中:$\boldsymbol{Y}$是$n\times1$维响应变量;ρ是空间自相关系数,$\boldsymbol{W}$是行标准化后的空间权重矩阵;$\boldsymbol{X}(t)$是紧集Γ上的二阶随机过程且满足$E(\boldsymbol{X}(t))=0$、$E(\int_{\Gamma}\boldsymbol{X}^2(t)\mathrm{d}t)<\infty$;$\beta(t)$是未知的斜率函数;$\boldsymbol{\varepsilon}$是随机误差且与$\boldsymbol{X}(t)$相互独立。

SP-SARM的一般形式[16]为

$$\begin{cases}\boldsymbol{Y}=\rho\boldsymbol{W}\boldsymbol{Y}+\boldsymbol{X}\boldsymbol{\beta}+g(\boldsymbol{Z})+\boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon}\overset{iid}{\sim}N(0,\sigma^2\boldsymbol{I}_n),\end{cases}\tag{1.4}$$

式中:$\boldsymbol{Y}$是$n\times1$维响应变量;ρ是空间自相关系数;$\boldsymbol{W}$是行标准化后的空间权重矩阵;$\boldsymbol{X}$是$n\times k$维自变量;$\boldsymbol{\beta}$是$k\times1$维回归系数向量;$\boldsymbol{Z}$是$n\times1$维协变量;$g(\cdot)$是未知的连接函数;$\boldsymbol{\varepsilon}$是随机误差项。

1.2 半参数函数型线性空间自回归模型的构建

式(1.4)中的SP-SARM只是针对自变量$\boldsymbol{X}$和协变量$\boldsymbol{Z}$都是普通矩阵型数据的情形,且$\boldsymbol{X}$和$\boldsymbol{Z}$都和地理位置有关[17];式(1.2)中的FLM适用于自变量$\boldsymbol{X}(t)$为函数型数据的情形。而SP-SARM中的自变量有时可取函数型数据,这时可考虑将式(1.4)的SP-SARM与式(1.2)的FLM结合,得到SP-FLSARM,具体建模过程如下:

假设观测到函数型自变量样本$x_i(t)$、协变量样本z_i和响应变量样本y_i,样本量为n;$x_i(t)$是来自于$\boldsymbol{X}(t)$的独立同分布样本,$\boldsymbol{X}(t)$是紧集Γ上平方可积的二阶矩随机过程且满足$E(\boldsymbol{X}(t))=0$、$E(\int_{\Gamma}\boldsymbol{X}^2(t)\mathrm{d}t)<\infty$[18]。不失一般性,假设$t\in[0,1]$,则SP-FLSARM为

$$\begin{cases}\boldsymbol{Y}=\rho\boldsymbol{W}\boldsymbol{Y}+\int_0^1\boldsymbol{X}(t)\beta(t)\mathrm{d}t+g(\boldsymbol{Z})+\boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon}\overset{iid}{\sim}N(0,\sigma^2\boldsymbol{I}_n),\end{cases}\tag{1.5}$$

式中:$\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$是响应变量;$\rho$是空间自相关系数且$\rho\in[0,1]$;$\boldsymbol{W}$是行标准化后的空间权重矩阵;$\boldsymbol{X}(t)=(x_1(t),x_2(t),\cdots,x_n(t))^{\mathrm{T}}$是函数型自变量;$\beta(t)$是定义在$[0,1]$上平方可积的斜率函数;$\boldsymbol{Z}=(z_1,z_2,\cdots,z_n)^{\mathrm{T}}$是协变量;$g(\cdot)$是未知的连接函数;$\boldsymbol{\varepsilon}$是随机误差项。

当SP-FLSARM表示为式(1.6)时,

$$\boldsymbol{Y}=(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\int_0^1\boldsymbol{X}(t)\beta(t)\mathrm{d}t+(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}g(\boldsymbol{Z})+(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon},\tag{1.6}$$

扰动项变成$(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon}$,显然扰动项之间不再独立,这与Gauss-Markov假设相矛盾,因此在参数估计时不能使用普通的最小二乘估计。

2 估计方法

由SP-FLSARM的模型式(1.5)可知,使用该模型时需要估计的参数有空间自相关系数ρ、斜率函数$\beta(t)$、连接函数$g(\cdot)$和随机误差的方差σ^2。本文利用FPCA处理函数型数据$\boldsymbol{X}(t)$,用B样条处理连接函数$g(\cdot)$,最后再利用MLE来估计SP-FLSARM

中的所有未知参数。

2.1　函数型数据的FPCA处理

记$K(s,t)$为$\boldsymbol{X}(t)$的协方差函数,即$K(s,t)=\mathrm{Cov}(\boldsymbol{X}(t),\boldsymbol{X}(s))$。利用Mercer定理,对协方差函数进行谱分解。即若k_j为特征值且$k_j>0(j=1,2,\cdots)$,$\{\varphi_j(t)\}_{j=1}^{\infty}$为相互正交的特征向量,则有

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

通过K-L变换,$\boldsymbol{X}(t)$可以展开为

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

式中$\boldsymbol{a}_j=\int_0^1\boldsymbol{X}(t)\varphi_j(t)\mathrm{d}t$是不相关的随机变量,其中$E(\boldsymbol{a}_j)=0$且$E(\boldsymbol{a}_j^2)=k_j$[11]。

$\boldsymbol{X}(t)$的样本协方差阵可以写为

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

式中$\hat{k}_j$和$\hat{\varphi}_j(t)$分别为k_j和$\varphi_j(t)$的估计值。

对于第i个观测$x_i(t)$,a_{ij}的估计值为$\hat{a}_{ij}=\int_0^1x_i(t)\hat{\varphi}_j(t)$,那么此时$x_i(t)$可以写为$x_i(t)=\sum_{j=1}^{\infty}\hat{a}_{ij}\hat{\varphi}_j(t)$。同理,$\beta(t)$可以写为$\beta(t)=\sum_{j=1}^{\infty}b_j\hat{\varphi}_j(t)$,其中$b_j=\int_0^1\beta(t)\hat{\varphi}_j(t)\mathrm{d}t$。此时,

$$\int_0^1\boldsymbol{X}(t)\beta(t)\mathrm{d}t=\sum_{j=1}^{\infty}\hat{\boldsymbol{a}}_jb_j,\qquad(2.1)$$

式中$\hat{\boldsymbol{a}}_j=(\hat{a}_{1j},\hat{a}_{2j},\cdots,\hat{a}_{nj})^{\mathrm{T}}$。

2.2　连接函数$g(\cdot)$的B样条(B-Spline)处理

对于一个非参数回归$Y_i=g(Z_i)+\varepsilon_i$而言,样条光滑回归利用了对分散点进行分段多项式拟合的思想,通过寻找一个满足条件的光滑函数使得残差平方和$\sum_{i=1}^{n}(y_i-g(z_i))^2$极小化[19]。

由B样条基函数构建的样条函数可写作

$$g(z)\approx\eta_1B_1(z)+\eta_2B_2(z)+\cdots+\eta_SB_S(z)=\boldsymbol{B}(\boldsymbol{Z})^{\mathrm{T}}\boldsymbol{\eta},\qquad(2.2)$$

式中:$\boldsymbol{B}(\boldsymbol{Z})=[B_1(z),B_2(z),\cdots,B_S(z)]^{\mathrm{T}}$,$\boldsymbol{\eta}=[\eta_1,\eta_2,\cdots,\eta_S]^{\mathrm{T}}$。

2.3　截断的SP-FLSARM

经过式(2.1)的FPCA和式(2.2)的B样条处理,得到SP-FLSARM的估计式(I):

$$\boldsymbol{Y}\approx\rho\boldsymbol{W}\boldsymbol{Y}+\sum_{j=1}^{\infty}\hat{\boldsymbol{a}}_jb_j+\sum_{s=1}^{S}\eta_sB_s(z)+\boldsymbol{\varepsilon},\qquad(2.3)$$

利用经典的赤池信息和贝叶斯信息准则方法,选取FPCA中特征函数的个数和B样条中基函数的个数,这里选取当2种方法值达到最小时的前m个特征函数和前J个基函数,并忽略截断误差,得到模型式(2.4)的截断SP-FLSARM,即SP-FLSARM的估计式(II)

$$\boldsymbol{Y}\approx\rho\boldsymbol{W}\boldsymbol{Y}+\sum_{j=1}^{m}\hat{\boldsymbol{a}}_jb_j+\sum_{s=1}^{J}\eta_sB_s(z)+\boldsymbol{\varepsilon}。\qquad(2.4)$$

2.4　利用MLE估计未知参数

为方便对SP-FLSARM的估计式(II)中的未知参数进行估计,做以下符号说明:令$\boldsymbol{A}=(\hat{a}_{ij})_{n\times m}$,$\boldsymbol{b}=(b_1,b_2,\cdots,b_m)^{\mathrm{T}}$,$\boldsymbol{B}(\boldsymbol{Z})$与$[B_1(z),B_2(z),\cdots,B_J(z)]^{\mathrm{T}}$,$\boldsymbol{\eta}=[\eta_1,\eta_2,\cdots,\eta_J]^{\mathrm{T}}$,可得到SP-FLSARM的估计式

$$\begin{aligned}\boldsymbol{Y}&\approx(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}[\boldsymbol{A}\boldsymbol{b}+\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})\boldsymbol{\eta}]+(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon}=\\&(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})][\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\eta}^{\mathrm{T}}]^{\mathrm{T}}+(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon},\end{aligned}\qquad(2.5)$$

其中,$\boldsymbol{\varepsilon}\sim N[0,\sigma^2\boldsymbol{I}_n]$。根据式(2.5)易知响应变量$\boldsymbol{Y}$的均值和协方差阵分别为:

$$\boldsymbol{\mu}=(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})][\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\eta}^{\mathrm{T}}]^{\mathrm{T}},$$

$$\boldsymbol{\Sigma}=[(\boldsymbol{I}_n-\rho\boldsymbol{W}^{\mathrm{T}})(\boldsymbol{I}_n-\rho\boldsymbol{W})]^{-1}\sigma^2。$$

即$\boldsymbol{Y}\sim N[\boldsymbol{\mu},\boldsymbol{\Sigma}]$。根据式(2.5),可知待估参数分别为$\rho$、$\boldsymbol{b}$、$\boldsymbol{\eta}$、$\sigma^2$。下面利用MLE估计未知参数,为了下文方便记叙,令$\boldsymbol{\Sigma}\equiv\sigma^2\boldsymbol{P}$,$\boldsymbol{R}\equiv\boldsymbol{P}^{-1}$,则可以写出对数似然函数

$$\begin{aligned}l(\theta)=&-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)+\frac{1}{2}\ln|\boldsymbol{R}|-\\&\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{Y}-\boldsymbol{\mu})。\end{aligned}\qquad(2.6)$$

下面分别对$l(\theta)$关于$(\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}$、$\sigma^2$求偏导,并令其为0,即

$$\begin{cases}\dfrac{\partial l(\theta)}{\partial(\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}}=0,\\[2ex]\dfrac{\partial l(\theta)}{\partial\sigma^2}=0。\end{cases}$$

根据上式,可得到$(\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}$、$\sigma^2$的估计形式:

$$\begin{aligned}&[\hat{\boldsymbol{b}}^{\mathrm{T}}(\rho),\hat{\boldsymbol{\eta}}^{\mathrm{T}}(\rho)]^{\mathrm{T}}=\\&\{[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})]^{\mathrm{T}}[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})]\}^{-1}[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})]^{\mathrm{T}}(\boldsymbol{I}_n-\rho\boldsymbol{W})\boldsymbol{Y},\\&\hat{\sigma}^2(\rho)=\frac{1}{n}\{\boldsymbol{Y}-(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})]\\&[\hat{\boldsymbol{b}}^{\mathrm{T}}(\rho),\hat{\boldsymbol{\eta}}^{\mathrm{T}}(\rho)]^{\mathrm{T}}\}^{\mathrm{T}}\boldsymbol{R}\{\boldsymbol{Y}-(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\\&[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})][\hat{\boldsymbol{b}}^{\mathrm{T}}(\rho),\hat{\boldsymbol{\eta}}^{\mathrm{T}}(\rho)]^{\mathrm{T}}\},\end{aligned}\qquad(2.7)$$

易看出,式(2.7)中的$[\hat{\boldsymbol{b}}^{\mathrm{T}}(\rho),\hat{\boldsymbol{\eta}}^{\mathrm{T}}(\rho)]^{\mathrm{T}}$、$\hat{\sigma}^2(\rho)$均与$\rho$有关,将$[\hat{\boldsymbol{b}}^{\mathrm{T}}(\rho),\hat{\boldsymbol{\eta}}^{\mathrm{T}}(\rho)]^{\mathrm{T}}$、$\hat{\sigma}^2(\rho)$带入式(2.6),则对数似然函数只与未知参数$\rho$有关,即

$$l(\rho)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\hat{\sigma}^2(\rho))+\ln|\boldsymbol{I}_n-\rho\boldsymbol{W}|-\frac{1}{2}。\tag{2.8}$$

对式(2.8)进行MLE估计,即令$\frac{\partial l(\rho)}{\partial\rho}=0$,则$\rho$的估计值为

$$\hat{\rho}=\underset{\rho}{\operatorname{argmax}}\ l(\rho)。\tag{2.9}$$

将式(2.9)中的$\hat{\rho}$代入到式(2.7)可得到:

$$\begin{aligned}[\hat{\boldsymbol{b}}^{\mathrm{T}},\hat{\boldsymbol{\eta}}^{\mathrm{T}}]^{\mathrm{T}}&=\{[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})]^{\mathrm{T}}[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})]\}^{-1}\\&\quad[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})]^{\mathrm{T}}(\boldsymbol{I}_n-\hat{\rho}\boldsymbol{W})\boldsymbol{Y},\\\hat{\sigma}^2&=\frac{1}{n}\{\boldsymbol{Y}-(\boldsymbol{I}_n-\hat{\rho}\boldsymbol{W})^{-1}[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})][\hat{\boldsymbol{b}}^{\mathrm{T}},\hat{\boldsymbol{\eta}}^{\mathrm{T}}]^{\mathrm{T}}\}^{\mathrm{T}}\boldsymbol{R}\\&\quad\{\boldsymbol{Y}-(\boldsymbol{I}_n-\hat{\rho}\boldsymbol{W})^{-1}[\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z})][\hat{\boldsymbol{b}}^{\mathrm{T}},\hat{\boldsymbol{\eta}}^{\mathrm{T}}]^{\mathrm{T}}\},\end{aligned}\tag{2.10}$$

于是,$\beta(t)$的估计值$\hat{\beta}(t)$可以表示为

$$\hat{\beta}(t)=\sum_{j=1}^{m}\hat{b}_j\hat{\varphi}_j(t),\tag{2.11}$$

$$\hat{g}(.)=\boldsymbol{B}(.)^{\mathrm{T}}\hat{\boldsymbol{\eta}}。\tag{2.12}$$

式中$\hat{b}_j$和$\hat{\boldsymbol{\eta}}$分别从$[\hat{\boldsymbol{b}}^{\mathrm{T}},\hat{\boldsymbol{\eta}}^{\mathrm{T}}]^{\mathrm{T}}$中获取。

3 模拟分析

为了验证上文构建的SP-FLSARM估计式(III)中参数估计的有效性,下面对上文中提出的基于FPCA和B样条的MLE方法进行模拟分析。使用Rstudio软件生成符合SP-FLSARM要求的数据,并调用spdep、fda和fda.usc包,利用文中的估计方法对生成的数据进行模拟分析,分别对比当ρ=0.3、0.5、0.7时$\hat{\rho}$、$\hat{\sigma}^2$、$\hat{\beta}(t)$、$\hat{g}(.)$的变化情况,据此评判上文提出的估计方法的优劣性。

3.1 数据的产生

在模拟实验中生成$\boldsymbol{x}(t)$时,样本大小记作n=$R\times M$。根据Hall和Horowitz[20]在FLM中提出的方法,函数型自变量$\boldsymbol{x}(t)=(x_1(t),x_2(t),\cdots,x_n(t))^{\mathrm{T}}$,以及斜率函数$\beta(t)$和连接函数$g(\cdot)$可分别独立地产生,

$$x_i(t)=\sum_{j=1}^{10}a_jV_j\varphi_j(t),i=1,2,\cdots,n,\tag{3.1}$$

式中:$a_j=(-1)^{j+1}j^{-\gamma/2}$,$\gamma$取1.1;$V_j\sim U[-\sqrt{3},\sqrt{3}]$;$\varphi_j(t)=\sqrt{2}\cos(j\pi t)$。

$$\beta(t)=\sum_{j=1}^{10}b_j\varphi_j(t),\tag{3.2}$$

式中:$b_1=0.3$;$b_j=4(-1)^{j+1}j^{-2},j\geqslant 2$。

$$g(z)=\mathrm{e}^{2z}\cos(2\pi z),\tag{3.3}$$

式中$\boldsymbol{z}=(z_1,z_2,\cdots,z_n)^{\mathrm{T}}$为协变量,其中$z_i\sim U[0,1]$,$i=1,2,\cdots,n$。

SP-FLSARM的响应变量数据由式(3.4)产生,

$$\begin{aligned}\boldsymbol{y}&=(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\left(\int_0^1\boldsymbol{x}(t)\beta(t)\mathrm{d}t+g(\boldsymbol{z})+\boldsymbol{\varepsilon}\right),\\\boldsymbol{\varepsilon}&\overset{iid}{\sim}N[0,\boldsymbol{I}_n]。\end{aligned}\tag{3.4}$$

3.2 SP-FLSARM的模拟结果

为了保证模拟实验的有效性,分别取样本量n=100、300、500,空间自相关系数ρ=0.3、0.5、0.7,针对n和ρ的每一组取值均进行500次模拟实验,并计算$\hat{\rho}$、$\hat{\sigma}^2$的平均值、中位数、偏差和标准差的结果(表1),并以此来评估ρ、σ^2的估计效果。显然,无论ρ取0.3、0.5还是0.7,$\hat{\rho}$和$\hat{\sigma}^2$的平均值、中位数都与真实值非常接近,且$\hat{\rho}$、$\hat{\sigma}^2$的标准差和偏差也都较小。同时,随着样本量n逐渐增大,ρ、σ^2的估计效果也逐渐变好。当样本量n=500时,ρ和σ^2的估计值最接近真实值,标准差和偏差也最小;当样本量n=300

表1 500次模拟实验中$\hat{\rho}$和$\hat{\sigma}^2$的相关计算结果

ρ	n	$\hat{\rho}$				$\hat{\sigma}^2$			
		均值	中位数	标准差	偏差	均值	中位数	标准差	偏差
0.3	100	0.294 7	0.295 4	0.003 8	−0.025 0	0.970 5	0.956 3	0.010 9	−0.031 0
	300	0.297 3	0.298 2	0.002 5	−0.009 7	0.988 2	0.980 6	0.003 9	−0.011 7
	500	0.300 5	0.301 3	0.000 4	−0.002 5	1.003 8	0.998 9	0.001 7	0.004 3
0.5	100	0.497 5	0.501 0	0.003 4	−0.012 4	1.063 7	1.050 0	0.008 2	0.063 7
	300	0.498 0	0.506 8	0.002 4	−0.010 9	1.027 2	1.027 0	0.003 7	0.027 2
	500	0.499 6	0.500 1	0.000 7	−0.002 3	1.000 4	1.000 1	0.001 3	0.003 7
0.7	100	0.694 6	0.700 4	0.003 1	−0.015 3	1.051 5	1.041 7	0.010 1	0.054 9
	300	0.702 1	0.703 8	0.002 6	0.008 6	1.031 1	1.029 1	0.003 3	0.026 9
	500	0.699 4	0.700 4	0.000 5	−0.003 1	1.001 8	1.001 3	0.001 5	0.004 0

注:ρ为空间自相关系数;n为样本量;$\hat{\rho}$为ρ的估计值;$\hat{\sigma}^2$为σ^2的估计值。

时,ρ和σ^2的估计效果次之;当样本量$n=100$时,ρ和σ^2的估计效果不如样本量为300时的结果。

然后,利用积分均方误差(integrated mean square error, IMSE)(E_{IMS})来评估$\beta(t)$和$g(\cdot)$的估计效果,E_{IMS}越小,$\beta(t)$和$g(\cdot)$的估计效果越好。将$\beta(t)$在[0,1]上分成300个等间距的点$\{t_k\}_{k=1}^{300}$,则

$$E_{\mathrm{IMS}}=\int(\hat{\beta}(t)-\beta(t))^2\,\mathrm{d}t\approx\frac{1}{300}\sum_{k=1}^{300}(\hat{\beta}(t_k)-\beta(t_k))^2,\tag{3.5}$$

式中$\hat{\beta}(t_k)$是$\beta(t)$在t_k处的估计值。同理,也可求$g(\cdot)$的E_{IMS}。接下来计算500次模拟实验中$\beta(t)$和$g(\cdot)$的E_{IMS}的平均值、中位数和标准差,计算结果如表2所示。当$n=100$、300、500时,$\beta(t)$和$g(\cdot)$的E_{IMS}的平均值、中位数和标准差都很小;随着样本量n逐渐增大,$\beta(t)$和$g(\cdot)$的E_{IMS}的平均值、中位数和标准差越来越小。这说明随着样本量n的增大,$\beta(t)$和$g(\boldsymbol{Z})$的估计效果越来越好。

为了更加直观地反映该估计方法的优劣性,在500次模拟结果中,选取E_{IMS}最小时对应的$\beta(t)$和$g(\boldsymbol{Z})$的估计值,分别绘制$\beta(t)$和$g(\boldsymbol{Z})$的曲线如图1和2所示。估计曲线和真实曲线都非常接近,这说明$\beta(t)$和$g(\boldsymbol{Z})$的估计值与真实值非常接近,这也表明上文中基于FPCA和B样条的MLE方法合理有效。同时,样本量n越大,2条曲线越相近,这也符合表1和2所呈现的数据规律。

综上,无论从模拟数据的比对分析,还是所选图形的直观比较,都表明基于FPCA和B样条的MLE方法能够精确地估计出SP-FLSARM中的未知参数,因此,可运用文中构建的SP-FLSARM进行实际数据的分析。

4　SP-FLSARM的实例分析

4.1　西班牙气象台数据

为了说明SP-FLSARM的实际应用价值,本文选取Rstudio的数据包aemet中的西班牙气象台的光照数据进行实例分析。选取1980—2009年西班牙73个气象台收集到的年平均日温度数据,关于年份求平均值,以此作为函数型自变量,年平均日温度曲线如图3所示。(1)函数型自变量的维数为73×365。大部分的曲线都能按照共同的特征聚合在一起,但其中8条曲线因为离群点导致脱离了这个群体,见图3(a)。为了保证模型结果的有效性,首先剔除8个离群点,得到65个气象台30年的年平均日温度曲线,所有曲线都能按照共同的特征聚合在一起,且每条曲线的波动趋势一致,见图3(b)。(2)协变量为73个气象台的海拔数据,其维数为73×1。(3)响应变量为73个气象台收集到的晴天数的年平均值。通过查看73个气象台的晴天数,响应变量中有14个缺失值,这里直接将其剔除。剔除离群点和缺失数据后,函数型自变量的维数为51×365,协变量的维数为51×1,响应变量的维数为51×1。

在处理空间地理数据时最重要的是判断地理位置间是否存在空间自相关性。空间自相关指各个单位、地点之间的相互依赖程度是由地理位置或经济相对位置决定的,一般情况下,认为距离越近,依赖程度越强[21]。空间效应的纳入需借助空间权重矩阵,考虑到该实例是关于温度、海拔和晴天数的分析,各个气象台站点获取的数据与气象台所处位置有一定的关系,因此,这里空间权重矩阵选用负指数权重矩阵,空间权重$\boldsymbol{W}_{ij}$的计算公式为

表2　500次模拟实验中$\beta(t)$和$g(Z)$的IMSE的相关计算结果

ρ	n	$\hat{\beta}(t)$			$\hat{g}(\boldsymbol{Z})$		
		E_{IMS}均值	E_{IMS}中位数	E_{IMS}标准差	E_{IMS}均值	E_{IMS}中位数	E_{IMS}标准差
0.3	100	0.313 3	0.265 9	0.005 4	0.071 6	0.065 2	0.001 80
	300	0.144 1	0.144 5	0.002 8	0.026 6	0.023 1	0.000 18
	500	0.108 9	0.102 6	0.000 9	0.010 3	0.012 7	0.000 04
0.5	100	0.312 0	0.271 6	0.004 9	0.073 1	0.065 3	0.001 70
	300	0.154 1	0.151 0	0.003 1	0.023 9	0.021 2	0.000 16
	500	0.100 3	0.101 0	0.001 5	0.011 4	0.012 4	0.000 08
0.7	100	0.315 1	0.294 6	0.005 1	0.074 7	0.068 5	0.002 00
	300	0.152 2	0.151 9	0.002 9	0.023 7	0.022 5	0.000 19
	500	0.101 3	0.101 7	0.001 1	0.012 1	0.011 5	0.000 06

注:$\beta(t)$为斜率函数;$g(\boldsymbol{Z})$为连接函数;$\hat{\beta}(t)$和$\hat{g}(\boldsymbol{Z})$分别为$\beta(t)$和$g(\boldsymbol{Z})$的估计值;E_{IMS}为积分均方误差。

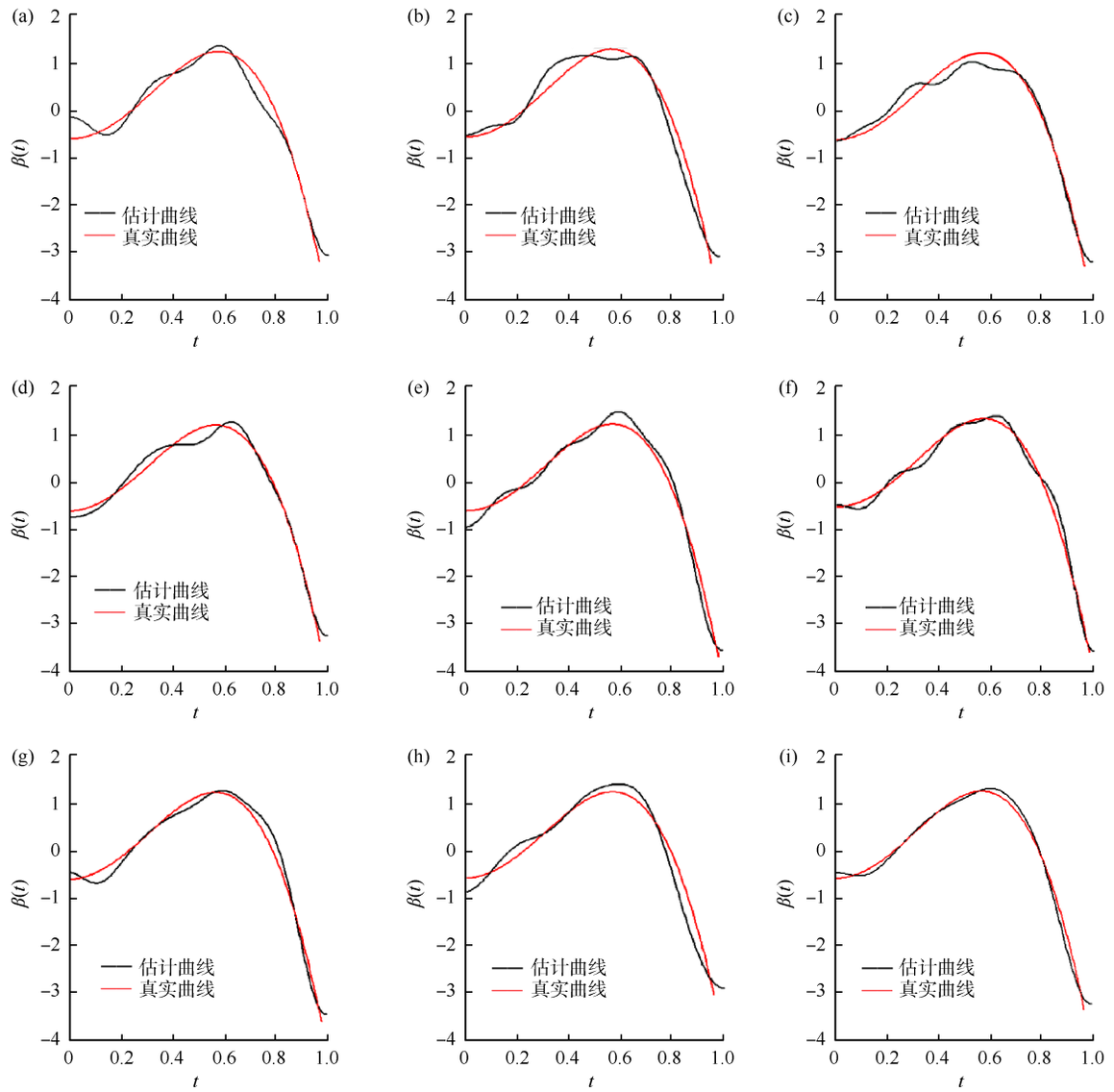


图1　不同n和ρ时$\beta(t)$的相应曲线

(a)n = 100,ρ = 0.3;(b)n = 100,ρ = 0.5;(c)n = 100,ρ = 0.7;(d)n = 300,ρ = 0.3;(e)n = 300,ρ = 0.5;(f)n = 300,ρ = 0.7;(g)n = 500,ρ = 0.3;(h)n = 500,ρ = 0.5;(i)n = 500,ρ = 0.7

$$\boldsymbol{W}_{ij}=\frac{\exp\left(-\left\|s_i-s_j\right\|\right)}{\sum_{l\neq i}\exp\left(-\left\|s_i-s_l\right\|\right)},$$

式中:s_i是第i个气象站的位置;$\|\cdot\|$代表欧氏距离。

判断空间自相关性需用到Moran's I检验,Moran's I=0.707,即本文使用的负指数权重矩阵通过了Moran's I检验,这说明剩余的51个气象台的空间自相关性是显著的,因此,可以利用SP-FLSARM对西班牙气象台的数据进行分析,并对模型中的相应参数进行估计。

4.2　利用SP-FLSARM分析西班牙气象台数据

为说明本文构建的SP-FLSARM的有效性和优越性,将SP-FLSARM、FLSARM和FLM共3种模型应用于西班牙气象台数据,得到对应的斜率函数$\beta(t)$和连接函数$g(\boldsymbol{Z})$的估计曲线如图4所示。

下面对SP-FLSARM、FLSARM、FLM模型中ρ、σ^2的估计值和模型的残差方差比进行分析。残差方差比(variance ratio,VR)的计算公式为

$$V_{\mathrm{R}}=\frac{\mathrm{var}(\boldsymbol{Y}-\hat{\boldsymbol{Y}})}{\mathrm{var}(\boldsymbol{Y})},$$

式中$\hat{\boldsymbol{Y}}$为响应变量的预测值。V_{R}越小,表明模型的拟合效果较好。3种模型的估计结果列于表3,与FLSARM和FLM相比,SP-FLSARM的$\hat{\sigma}^2$和V_{R}均最小,这说明SP-FLSARM的估计效果是最好的。另外,利用FLM处理西班牙气象台数据后,残差的

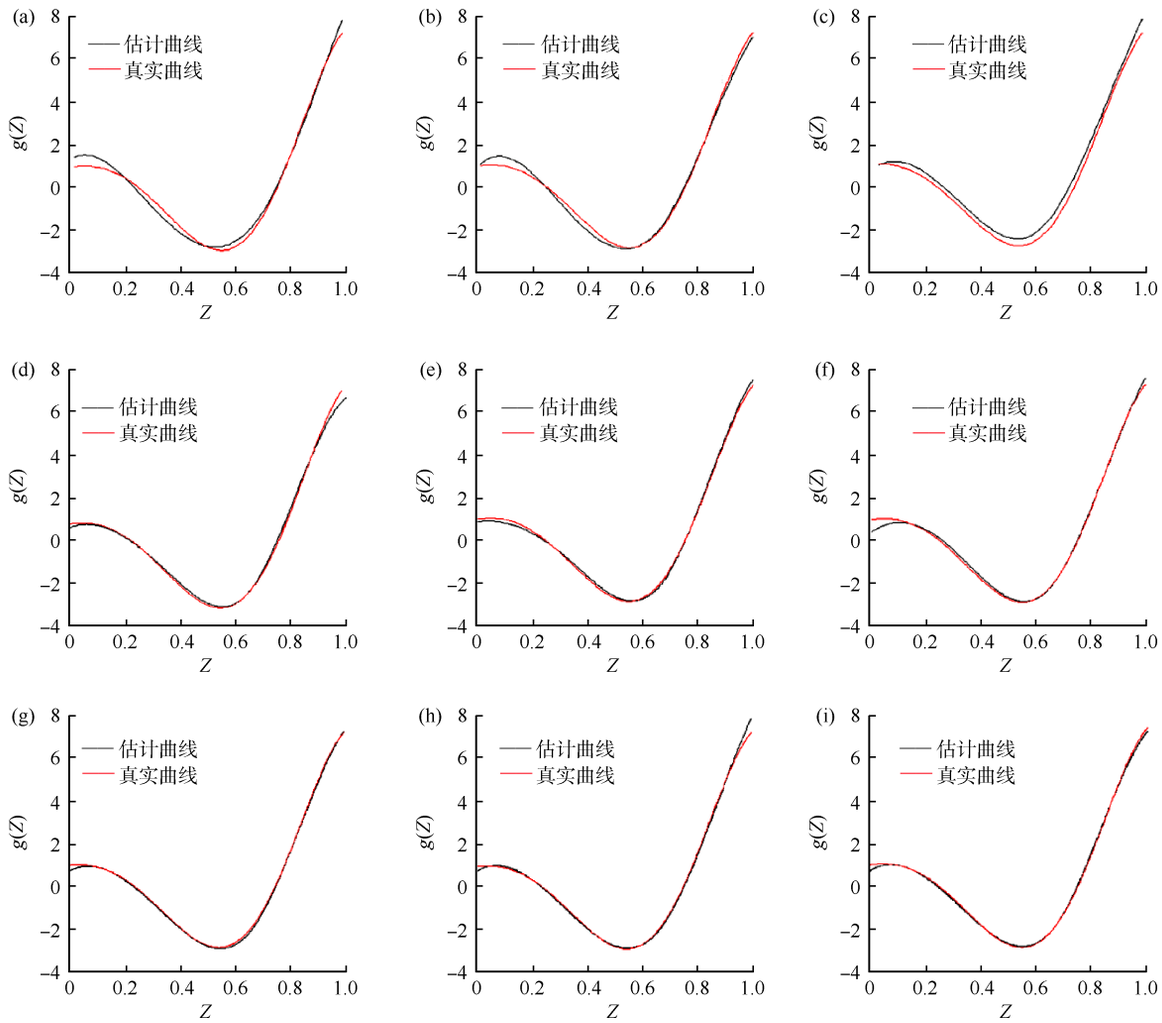


图 2　不同 n 和 ρ 时 $g(Z)$ 的相应曲线

(a)n = 100,ρ = 0.3;(b)n = 100,ρ = 0.5;(c)n = 100,ρ = 0.7;(d)n = 300,ρ = 0.3;(e)n = 300,ρ = 0.5;(f)n = 300,ρ = 0.7;(g)n = 500,ρ = 0.3;(h)n = 500,ρ = 0.5;(i)n = 500,ρ = 0.7

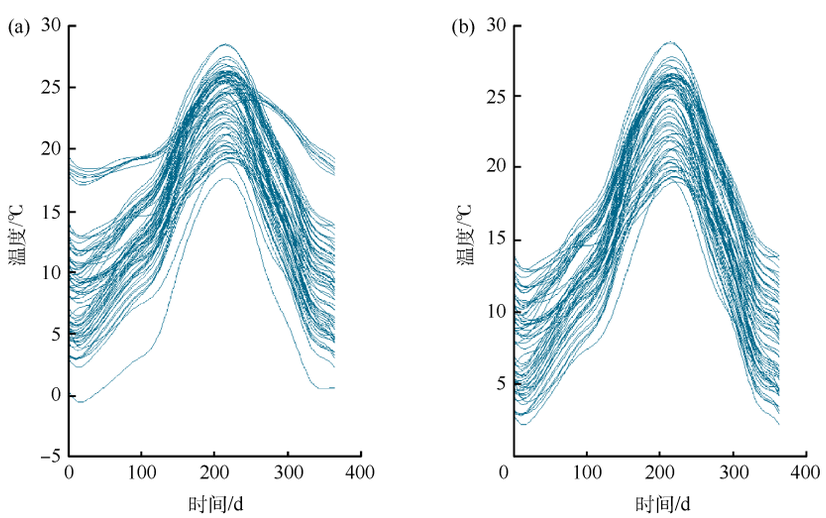
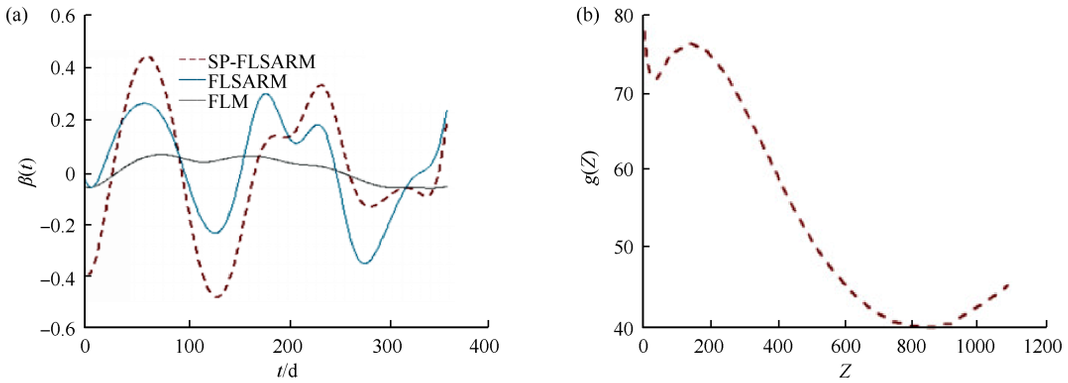


图 3　1980—2009 年西班牙气象台收集的年平均日温度曲线

(a)年平均日温度;(b)剔除离群点后的年平均日温度



注:SP-FLSARM为半参数函数型线性空间自回归模型;FLSARM为函数型线性空间自回归模型;FLM为函数型线性模型。

图4　3种模型的斜率函数$\beta(t)$和连接函数$g(Z)$的估计曲线

(a)3种模型的斜率函数$\beta(t)$的估计曲线;(b)SP-FLSARM中连接函数$g(Z)$的估计曲线

Moran's I值为0.213,P=0.002<0.05,即拒绝残差不存在空间效应的原假设,这说明利用FLM处理西班牙气象台数据时,空间效应提取不彻底,残差依然存在空间效应;而SP-FLSARM和FLSARM提取空间效应后,残差的Moran's I值分别为0.007和0.089,P=0.327和0.074,P > 0.05,均无法拒绝残差不存在空间效应的原假设,即提取过空间效应后,残差不存在空间效应。这也说明在处理具有空间效应的数据时,SP-FLSARM优于FLSARM和FLM。

根据SP-FLSARM、FLSARM和FLM模型对响应变量的数据进行预测,响应变量的预测值和真实值做出的散点图如图5所示。图中散点均在对角

表3　3种模型的估计结果

模型	$\hat{\rho}$	$\hat{\sigma}^2$	V_R	残差的Moran's I值	P
SP-FLSARM	0.698	91.070	0.110 9	0.007	0.327
FLSARM	0.684	170.296	0.126 2	0.089	0.074
FLM	—	214.824	0.144 5	0.213	0.002

注:V_R为残差方差比;P为显著性检验的值;—指无数据。

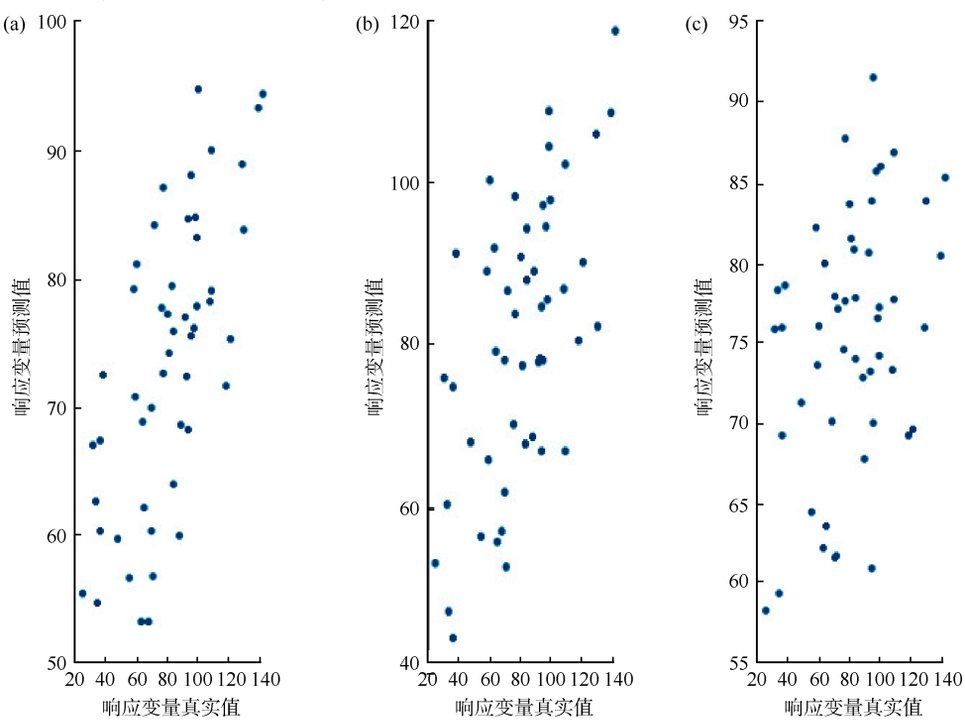


图5　3种模型响应变量的真实值和预测值的散点图

(a)SP-FLSARM;(b)FLSARM;(c)FLM

线附近上下波动,利用SP-FLSARM、FLSARM和FLM对西班牙51个气象台收集的晴天数数据进行预测时,效果都比较好。同时,基于SP-FLSARM的散点图相较于FLSARM和FLM,散点分布更集中,这也说明SP-FLSARM的预测效果最好。

5 结束语

本文提出了SP-FLSARM模型,并分别利用FPCA和B样条处理函数型自变量和非参数连接函数,再用MLE对未知参数进行估计,并将SP-FLSARM模型应用到西班牙气象台温度数据中,模型拟合效果良好。但在实际数据的收集过程中,总会不可避免地存在缺失数据,因此,对于SP-FLSARM模型而言,如何有效地处理缺失数据是至关重要的,这也是可继续深入研究的方向之一。

参 考 文 献

[1] TOBLER W R. A computer movie simulating urban growth in the Detroit Region[J]. Economic Geography, 1970,46: 234-240.

[2] CARDOT H, FERRATY F, SARDA P. Functional linear model[J]. Statistics Probability & Letters, 1999, 45(1): 11-22.

[3] MARZIO M D, PANZERA A, VENIERI C. Non-parametric regression for compositional data [J]. Statistical Modelling, 2015, 15(2): 113-133.

[4] AW A, CABRAL E N. Bayesian estimation of the functional spatial lag model[J]. Journal of Time Series Econometrics, 2020, 12(2): 1-11.

[5] ANSELIN L. Spatial econometrics: methods and models [J]. Economic Geography, 1988, 65(2): 160-162.

[6] FEBRERO-BANDE M, GALEANO P, GONAÁLEZ-MANTEIGA W. Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random[J]. Computational Statistics and Data Analysis, 2018, 12: 1-13.

[7] ZHANG L, BALADANDAYUTHAPANI V, ZHU H, et al. Functional CAR models for large spatially correlated functional datasets [J]. Journal American Statistical Association, 2016, 111(514): 772-786.

[8] PINDA-RÍOS W, GIRALDO R, PORCU E. Functional SAR models: with application to spatial econometrics [J]. Spatial Statistics, 2019(29): 145-159.

[9] AW A, CABRAL E N. Functional SAC model: with application to spatial econometrics [J]. Department of Mathematics, 2020, 16(4): 1-12.

[10] 胡锡健. 部分函数型空间回归模型及其应用研究[D]. 乌鲁木齐: 新疆大学, 2021: 1-52.

[11] BAÍLLO A, GRANÉ A. Local linear regression for functional predictor and scalar response [J]. Journal of Multivariate Analysis, 2009, 100(1): 102-111.

[12] WANG H, LIN J, WANG J. Nonparametric spatial regression with spatial autoregressive error structure[J]. Statistics, 2015, 50(1): 60-75.

[13] SU L, JIN S. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models [J]. Journal of Econometrics, 2010, 157(1): 18-33.

[14] 李序颖, 顾岚. 空间自回归模型及其估计[J]. 统计研究, 2004(6): 48-51.

[15] PAGANONI A M, SANGALLI L M. Functional regression models: some directions of future research [J]. Statistical Modelling, 2017, 17(1/2): 94-99.

[16] 程素丽. 半参数空间回归模型的GMM估计[D]. 福州: 福建师范大学, 2019:13-46.

[17] 尤培颖. 带随机缺失的半参数空间自回归模型的估计[D]. 昆明: 云南大学, 2018:1-52.

[18] 汤杨冰. 半参数空间自回归模型的秩估计[D]. 北京: 北京工业大学, 2020: 1-38.

[19] 马晓跃, 武新乾. 非参数回归模型基于残差的样条估计[J]. 河南科技大学学报(自然科学版), 2021, 42(4): 91-96+10.

[20] HALL P, HOROWITZ J L. Methodology and convergence rates for functional linear regression [J]. The Annals of Statistics, 2007, 35(1): 70-91.

[21] 郭强, 郑贵省, 马文彬. 基于连通度的空间权重矩阵构建方法[J]. 军事交通学院学报, 2021, 23(3): 92-95.

(责任编辑:马田田)