DOI:10.11830/ISSN.1000-5013.202108037

采用最小二乘支持向量机的部分相依函数型线性模型估计与应用

苏枙芳[1],周煜[1],李气芳[2]

(1. 华侨大学 经济与金融学院,福建 泉州 362021;
2. 闽南师范大学 数学与统计学院,福建 漳州 363000)

摘要: 提出一种基于无截断 Bartlett 核函数的重构方法,有效避免长期方差函数估计方法面临的核函数与窗宽选择问题,并将其应用到部分相依函数型线性模型中.利用考虑函数型数据相依性的最小二乘支持向量机对模型进行参数估计,数值模拟结果表明:与未考虑函数型数据相依特征的最小二乘估计方法相比,提出的考虑函数型数据相依性的最小二乘支持向量机估计方法能更稳健地估计向量系数,有效提高样本外的预测精度;将部分相依函数型线性模型应用到上证指数开盘价的预测中,得到较好的预测效果.

关键词: 部分相依函数型线性模型;长期协方差函数;相依函数型数据;最小二乘支持向量机

中图分类号: O 212 **文献标志码:** A **文章编号:** 1000-5013(2022)04-0544-09

Estimate and Application of Partial Dependent Functional Linear Model Using Least Squares Support Vector Machine

SU Zhifang[1], ZHOU Yu[1], LI Qifang[2]

(1. School of Economics and Finance, Huaqiao University, Quanzhou 362021, China;
2. School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, China)

Abstract: We propose the reconstruction method based on non truncated Bartlett kernel function, in which the selection of kernel function and window width faced by the long-term variance function estimation method are avoided effectively, and apply it to the partial dependent function linear model. The least squares support vector machine considering the dependence of function data is used to estimate the parameters of the model. Numerical simulation results show that, compared with the least squares estimation method not considering the dependent features of functional data, the least squares support vector machine estimation method considering the dependence of functional data is more robustand effectively improve the out-of-sample prediction accuracy. The partial dependent function linear model is applied to the prediction of the opening price of Shanghai stock index and a better prediction effect is obtained.

Keywords: partial dependent function linear model; long-term variance function; dependent functional datasleast; least squares support vector machine

随着数据采集、处理和存储技术的快速发展,越来越多的数据可被连续观测且在本质上呈现出明显的函数曲线特征,Ramsay 等[1]将这类数据定义为函数型数据,函数型数据分析已经广泛应用到气象

收稿日期: 2021-08-30

通信作者: 苏枙芳(1977-),男,教授,博士,博士生导师,主要从事数量经济模型、函数型数据分析方法的研究. E-mail:suzufine@hqu.edu.cn.

基金项目: 国家社科基金资助项目(21AJY001)

学、生物学、经济学等领域[2-5].

函数型线性模型是函数型数据分析的重要工具，Cardot 等[6-7]基于函数型主成分分析和惩罚样条的估计方法研究估计量的相关渐进性质. Yao 等[8]考虑观测值为稀疏离散情况下的函数型线性模型的估计方式. 文献[9-11]采用平滑样条方法估计函数型斜率参数，研究估计量的大样本性质.

为进一步提高函数型线性模型的预测能力和可解释性，Zhang 等[12]将向量型解释变量引入函数型线性模型中，提出部分函数型线性模型. Shin[13]运用函数主成分分析方法估计模型，并证明参数估计量的渐进正态性和函数系数估计量的最优收敛速度. Zhou 等[14]将模型的函数系数利用样条基展开，进一步通过最小二乘法得到估计量. 王晓光等[15]基于核函数构造一类部分函数线性回归模型，研究模型参数的渐进正态性和非参数的收敛速度.

现有的这些估计方法一般都假设函数型数据服从独立同分布(i. i. d)，而没有考虑函数型数据的相依特征. 现实生活中，股票数据、温度数据、空气污染物数据等函数型数据明显存在相依结构，如果运用独立同分布条件下的函数型数据分析方法重构这些数据，必然会出现误差，从而对后续模型的估计造成影响. 对此，文献[16-18]利用长期协方差函数替代独立同分布条件下的协方差函数，证明长期协方差函数收敛于总体长期协方差函数. 然而，长期协方差函数的估计涉及核函数和窗宽的选择易受人为因素的影响. 李气芳[19]在文献[20]的研究基础上，提出基于无截断 Bartlett 核的长期协方差函数估计方法，避免了核函数和窗宽的误选导致的估计误差. 综上，本文针对具有相依特征的函数型自变量，将独立同分布条件下的部分函数型线性模型拓展到相依情形.

1　模型与估计

1.1　部分相依函数型线性回归模型

针对自变量中同时含有标量型和函数型变量的情况，Zhang[9]提出了部分函数型线性回归模型，即观测数据$\{(X_1(t),Y_1,\boldsymbol{Z}_1),(X_2(t),Y_2,\boldsymbol{Z}_2),\cdots,(X_n(t),Y_n,\boldsymbol{Z}_n)\}$满足如下形式，即

$$Y_i=\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i+\int_0^1\beta(t)X_i(t)\mathrm{d}t+\varepsilon_i,\qquad i=1,2,\cdots,n. \tag{1}$$

式(1)中：$X_i(t)$为函数型变量，是 $L^2[0,1]$中的随机过程；$\beta(t)$为回归系数函数；$\boldsymbol{Z}_i$ 为 p 维标量型自变量；$\boldsymbol{\gamma}$ 为 p 维回归系数向量；ε_i 表示均值为 0，方差为 σ^2 的随机误差项，且与$(\boldsymbol{Z}_i,X_i(t))$独立；$Y_i$ 为标量型应变量.

若函数型数据 $X_i(t)$满足函数

$$\mathrm{Cov}[X_i(t),X_{i+h}(s)]=E\{[X_i(t)-\mu(t)][X_{i+h}(s)-\mu(s)]\}\neq 0,\qquad h\neq 0,$$

则称 $X_i(t)$为相依函数型数据. 当 $X_i(t)$为相依函数型数据时，可以把式(1)推广为部分相依函数型线性回归模型.

1.2　相依函数型数据的重构

函数型数据分析的首要任务是把函数型数据重构成函数曲线，其主要方法有外生基法(Fourier 基，B-Spline 基等)和内生基法(函数主成分基)，越来越多学者青睐函数主成分基的重构方法. 在独立同分布条件下，通过计算协方差函数得到函数主成分，但当函数型数据具有相依特征时，样本协方差函数不再是总体协方差函数的一致估计量，计算得到的函数主成分不准确. Hörmann 等[18]基于长期协方差函数计算函数主成分的方法，面临核函数和窗宽的选择问题. Kiefer 等[19]在研究多元回归模型中长期协方差估计问题时，构造基于无截断 Bartlett 核的长期协方差估计统计量，不需要选择核函数和窗宽. 李气芳[19]把文献[20]的估计思想推广到长期协方差函数的估计中. 因此，采用基于无截断 Bartlett 核的长期协方差函数估计方法，避免核函数和窗宽的选择问题.

相依函数型数据 $X_i(t)$的长期协方差函数定义为 $C(t,s)=\sum\limits_{h=-\infty}^{\infty}\mathrm{Cov}(X_0(t),X_h(s))$. 使用核函数对长期协方差函数进行估计，即

$$\hat{C}(t,s)=\hat{C}_0(t,s)+\sum_{h=1}^{n-1}K\left(\frac{h}{q}\right)[\hat{C}_h(t,s)+\hat{C}_{-h}(t,s)]. \tag{2}$$

式(2)中:$\hat{C}_0(t,s)=\frac{1}{n}\sum_{i=1}^{n}[X_i(t)-\overline{X}_n(t)][X_i(s)-\overline{X}_n(s)]$;$\hat{C}_h(t,s)=\frac{1}{n-h}\sum_{i=1}^{n-h}[X_i(t)-\overline{X}_n(t)][X_{i+h}(s)-\overline{X}_n(s)]$.

定义在$\hat{C}$中的核函数K需满足$K(0)=1$,$K(-u)=K(u)$,且当$|x|>1$时,$K(x)=0$.带宽q需满足n,$q\to\infty$时,$\frac{n}{q}=0$.

借鉴文献[19]中基于无截断Bartlett核的估计方法,把式(2)变为

$$\hat{C}(t,s)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(1-\frac{|i-j|}{n}\right)[X_i(t)-\overline{X}_n(t)][X_{i+h}(s)-\overline{X}_n(s)].$$

根据文献[21-22]对动态函数型主成分的定义,样本长期协方差函数的特征值与特征函数满足

$$\hat{\lambda}_i\hat{\varphi}_i(t)=\int_0^1\hat{C}(t,s)\hat{\varphi}_i(s)\mathrm{d}s,\qquad 1\leqslant i\leqslant n.$$

因此,特征值$\hat{\lambda}_1>\hat{\lambda}_2>\cdots>\hat{\lambda}_n>0$,特征函数序列$(\hat{\varphi}_1,\hat{\varphi}_2,\cdots,\hat{\varphi}_n)$是构成$L^2[0,1]$的一组正交基.

通过确定累积方差解释比例δ,$\sum_{i=1}^{m}\hat{\lambda}_i/\sum_{i=1}^{n-1}\hat{\lambda}_i>\delta$,决定主成分个数为$m$,进而求得$X_i(t)$第$k$个主成分得分,即

$$\hat{\xi}_{i,k}=\int_0^1(X_i(t)-\hat{\mu}(t))\hat{\varphi}_k(t)\mathrm{d}t. \tag{3}$$

基于Karhunen-Loeve展开,使用前m个函数主成分重构相依函数型数据,以达到降维的目的,即

$$\hat{X}_i(t)=\mu(t)+\sum_{k=1}^{\infty}\xi_{i,k}\varphi_k(t)\approx\hat{\mu}(t)+\sum_{k=1}^{m}\hat{\xi}_{i,k}\hat{\varphi}_k(t). \tag{4}$$

1.3 基于最小二乘支持向量机估计方法的系数估计

由式(4)得到的m个函数主成分对回归系数函数$\beta(t)$进行逼近,有

$$\beta(t)=\sum_{k=1}^{m}a_k\hat{\varphi}_k(t). \tag{5}$$

把式(4),(5)代入部分相依函数型线性模型,即

$$Y_i=\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i+\int_0^1\beta(t)X_i(t)\mathrm{d}t+\varepsilon_i,\qquad i=1,2,\cdots,n.$$

则有

$$Y_i=\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i+\sum_{k=1}^{m}a_k\langle\hat{\varphi}_k(t),\hat{X}_i(t)\rangle+\varepsilon_i=\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i+\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\eta}_i+\varepsilon_i,\qquad i=1,2,\cdots,n.$$

上式中:$\boldsymbol{a}^{\mathrm{T}}=(a_1\quad a_2\quad\cdots\quad a_m)^{\mathrm{T}}$;$\boldsymbol{\eta}_i=[\langle\hat{\varphi}_k(t),\hat{X}_i(t)\rangle]_{m\times1}$,$\boldsymbol{Z}_i=(Z_1\quad Z_2\quad Z_3\quad Z_4\quad Z_5)^{\mathrm{T}}$.

定义如下函数

$$l(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{a}})=\sum_{i=1}^{n}(Y_i-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i-\sum_{k=1}^{m}a_k\langle\hat{\varphi}_k(t),\hat{X}_i(t)\rangle)^2. \tag{6}$$

令$\boldsymbol{Y}=(Y_1\quad Y_2\quad\cdots\quad Y_n)^{\mathrm{T}}$,$\boldsymbol{A}=(\gamma_1\quad\cdots\quad\gamma_p\quad a_1\quad a_2\quad\cdots\quad a_m)^{\mathrm{T}}$,

$$\boldsymbol{B}=\begin{bmatrix}Z_{11}&\cdots&Z_{1p}&\langle\hat{\varphi}_1(t),\hat{X}_1(t)\rangle&\langle\hat{\varphi}_2(t),\hat{X}_1(t)\rangle&\cdots&\langle\hat{\varphi}_m(t),\hat{X}_1(t)\rangle\\Z_{21}&\cdots&Z_{2p}&\langle\hat{\varphi}_1(t),\hat{X}_2(t)\rangle&\langle\hat{\varphi}_2(t),\hat{X}_2(t)\rangle&\cdots&\langle\hat{\varphi}_m(t),\hat{X}_2(t)\rangle\\\vdots&&\vdots&\vdots&\vdots&&\vdots\\Z_{n1}&\cdots&Z_{np}&\langle\hat{\varphi}_1(t),\hat{X}_n(t)\rangle&\langle\hat{\varphi}_2(t),\hat{X}_n(t)\rangle&\cdots&\langle\hat{\varphi}_m(t),\hat{X}_n(t)\rangle\end{bmatrix}.$$

那么,式(6)可以改成线性回归模型的形式,即

$$l(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{a}})=(\boldsymbol{Y}-\boldsymbol{AB})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{AB}).$$

根据最小二乘法估计式,可得

$$(\hat{\boldsymbol{\gamma}}_{\mathrm{OLS}},\hat{\boldsymbol{a}}_{\mathrm{OLS}})=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}. \tag{7}$$

最小二乘法对样本容量要求较大且对异常值较敏感,而支持向量机算法引入了损失函数,允许一些

样本点出错，寻找的超平面只由少量支持向量决定，具有良好的鲁棒性. 最小二乘支持向量机估计方法是基于平方损失构建的一种支持向量机，其回归问题最终归结为等式约束下的线性方程组的求解问题，降低了计算的复杂度. 因此，运用最小二乘支持向量机算法，构造如下优化问题，即

$$\begin{cases}\min\limits_{\boldsymbol{\gamma},a,\beta} & \dfrac{1}{n}\sum\limits_{i=1}^{n}(Y_i-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\eta}_i)^2+\lambda\|\beta\|^2,\\ \text{s.t.} & Y_i=\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i+\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\eta}_i+\varepsilon_i, \qquad i=1,2,\cdots,n.\end{cases}$$

上式中：λ 为光滑参数. 将 $\lambda\|\beta\|^2$ 改写为

$$\lambda\|\beta\|^2=\lambda\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}\boldsymbol{a}.$$

上式中：$\widetilde{\boldsymbol{R}}=[\langle\hat{\varphi}_k,\hat{\varphi}_{k'}\rangle]_{k=1,k'=1}^{m}$.

引入拉格朗日乘子 μ_i，构建如下方程，即

$$L(\boldsymbol{\varepsilon},\boldsymbol{\gamma},\boldsymbol{a},\mu_i)=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2+\lambda\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}\boldsymbol{a}+\sum_{i=1}^{n}\mu_i(Y_i-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\eta}_i-\varepsilon_i).$$

分别对 $\boldsymbol{\varepsilon},\boldsymbol{\gamma},\boldsymbol{a},\mu_i$ 求偏导并等于零，可求得 $\hat{\boldsymbol{\gamma}}_{\mathrm{SVM}},\hat{\boldsymbol{a}}_{\mathrm{SVM}}$，即

$$\begin{bmatrix}\hat{\boldsymbol{\gamma}}_{\mathrm{SVM}}\\ \hat{\boldsymbol{a}}_{\mathrm{SVM}}\end{bmatrix}=\begin{bmatrix}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} & \boldsymbol{Z}\boldsymbol{\eta}^{\mathrm{T}}\\ \boldsymbol{\eta}\boldsymbol{Z}^{\mathrm{T}} & \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}+n\lambda\widetilde{\boldsymbol{R}}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{Z}\boldsymbol{Y}\\ \boldsymbol{\eta}\boldsymbol{Y}\end{bmatrix}. \tag{8}$$

将 $\hat{\boldsymbol{\gamma}}_{\mathrm{SVM}},\hat{\boldsymbol{a}}_{\mathrm{SVM}}$ 代入下列回归方程，可求得预测值 $\hat{\boldsymbol{Y}}$，即

$$\hat{\boldsymbol{Y}}=\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{Z}+\langle\hat{\beta},X\rangle. \tag{9}$$

2　数值模拟

2.1　模拟数据生成

样本数据由如下模型生成，即

$$Y_i=\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i+\int_0^1\beta(t)X_i(t)\mathrm{d}t+\varepsilon_i, \qquad i=1,2,\cdots,n.$$

上式中：系数向量 $\boldsymbol{\gamma}=(2.0\quad -1.0\quad 1.5\quad 5.0\quad -1.7)^{\mathrm{T}}$，随机向量 $\boldsymbol{Z}_i=(Z_1\quad Z_2\quad Z_3\quad Z_4\quad Z_5)^{\mathrm{T}}$，其与 $N(0,I_5)$ 同分布；随机误差 $\varepsilon_i\sim N(0,0.5^2)$.

对于相依函数型数据部分生成方法与文献[23]相同，相依函数型数据 $X_i(t)$ 服从 FAR(1) 过程，$X_i(t)=\int_0^1\varphi(t,s)X_{i-1}(t)\mathrm{d}s+\varepsilon_i(t)$，随机误差 $\varepsilon_i\sim N(0,1)$，$\varphi(t,s)=\dfrac{9}{4}\text{s.t.}\ X_i(t)$ 在 $[0,1]$ 上等间隔取 100 个观测点.

回归系数函数 $\beta(t)$ 有如下 3 个情形.

情形(Ⅰ)：$\beta(t)=0$.

情形(Ⅱ)：$\beta(t)=\sqrt{2}\sin(\pi t/2)+3\sqrt{2}\sin(3\pi t/2)$.

情形(Ⅲ)：$\beta(t)=\sum\limits_{j=1}^{50}b_j\varphi_j$，$b_1=0.5$，$\varphi_1=1$，$\varphi_j=\sqrt{2}\cos[(j-1)\pi t]$，$b_j=4j^{-2}$，$j\geqslant 2$.

2.2　模型参数估计的算法

模型参数估计的算法有如下 7 个步骤.

步骤 1　由 FAR(1) 过程 $X_i(t)=\int_0^1\varphi(t,s)X_{i-1}(t)\mathrm{d}s+\varepsilon_i$，生成 n 个相依函数型自变量.

步骤 2　由给定的 $\boldsymbol{\gamma},\beta(t),\boldsymbol{Z}_i,X_i(t),\varepsilon_i$ 结合回归模型(1)生成应变量 Y_i，得到数据集，把后 $0.2n$ 个样本作为样本外预测集.

步骤 3　不考虑函数型数据的相依性，计算 i.i.d. 条件下的样本短期协方差函数，然后，利用函数主成分重构得到 $\hat{X}_{1,i}(t)$，最后，使用最小二乘法估计式(7)得到 $\hat{\gamma}_1,\hat{\beta}_1$.

步骤 4　考虑函数型数据的相依性，利用无截断 Bartlett 核计算样本长期协方差，通过函数型主成

分基于 Karhunen-Loeve 展开重构,得到$\hat{X}_{2,i}(t)$.

步骤 5 通过留一交叉验证(CV)选取平滑参数λ,有

$$\mathrm{CV}(\lambda)=\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^{(-i)}-Y_i)^2.$$

上式中:$\hat{Y}_i^{(-i)}$ 是删去第 i 个模型的预测值 $\hat{Y}_i$.

步骤 6 将 $\lambda,\hat{X}_{2,i}(t),\boldsymbol{Z}_i,Y_i$ 代入最小二乘支持向量机估计式(8),得到 $\hat{\gamma}_2,\hat{\beta}_2$.

步骤 7 将估计得到 $\hat{\gamma}_1,\hat{\beta}_1$ 与 $\hat{\gamma}_2,\hat{\beta}_2$ 分别代入式(9),得到样本外预测值 $\hat{Y}_{1,i},\hat{Y}_{2,i}$.

2.3 模拟结果分析

为了说明文中方法的估计效果,重复步骤 1~6 共 500 次,定义平均偏离平方和(RASE)、均值(Mean)与方差作为 $\hat{\beta}$ 的评价指标;定义均方预测误差平方根(RMSPE)作为 $\hat{Y}$ 的评价指标,有

$$\mathrm{RASE}(\hat{\beta})=\frac{1}{100}\sum_{i=1}^{100}(\hat{\beta}(t_i)-\beta(t_i))^2,$$

$$\mathrm{Mean}(\hat{\beta})=\frac{1}{500}\sum_{j=1}^{500}\mathrm{RASE}^j,$$

$$\mathrm{Var}(\hat{\beta})=\frac{1}{500}\sum_{j=1}^{500}(\mathrm{RASE}^j)^2-(\mathrm{Mean})^2,$$

$$\mathrm{RMSPE}(\hat{Y})=\frac{1}{500}\sum_{j=1}^{500}\Big[\sum_{i=1}^{n}(\hat{Y}_i^j-Y_i^j)^2\Big]^{1/2}.$$

样本量 $n=50,100,200,500$ 的回归系数函数 $\beta(t)$ 估计得到 $\hat{\boldsymbol{\gamma}}$ 的偏差与方差,分别如表 1~4 所示.表 1~4 中:LSSVM 为文中方法,即考虑函数型数据相依性的最小二乘支持向量机方法;OLS 为未考虑函数型数据相依性的最小二乘估计方法;Bias 为计算偏误;Var 为方差;n 为样本数.

由表 1~4 可知:两种估计方法的偏误与方差非常接近且随着样本量的增大而减小,这说明两种估计方法在 3 种情形下都能取得较好的效果且性能表现近似.

表 1 三种情形下 $\hat{\boldsymbol{\gamma}}$ 的估计偏差与方差($n=50$)

Tab. 1 Deviation and variance of $\hat{\boldsymbol{\gamma}}$ estimated in three situations ($n=50$)

情形	方法	$\hat{\gamma}_1/\times10^{-3}$		$\hat{\gamma}_2/\times10^{-3}$		$\hat{\gamma}_3/\times10^{-3}$		$\hat{\gamma}_4/\times10^{-3}$		$\hat{\gamma}_5/\times10^{-3}$	
		Bias	Var	Bias	Var	Bias	Var	Bias	Var	Bias	Var
情形(Ⅰ)	LSSVM	−12.52	7.87	−11.15	4.46	1.66	5.73	15.85	6.01	−14.45	5.14
	OLS	−12.98	8.42	−11.03	5.06	2.48	6.39	18.04	6.53	−15.58	5.42
情形(Ⅱ)	LSSVM	3.34	6.91	2.09	5.01	8.49	6.49	−4.68	5.48	20.13	5.35
	OLS	2.69	7.15	2.30	5.35	9.70	6.90	−2.44	5.43	23.35	5.50
情形(Ⅲ)	LSSVM	1.79	5.98	−1.59	6.05	3.01	6.35	−12.70	6.73	5.77	4.70
	OLS	2.63	6.38	−2.66	6.54	6.66	7.45	−14.73	7.13	4.72	5.36

表 2 三种情形下 $\hat{\boldsymbol{\gamma}}$ 的偏差与方差($n=100$)

Tab. 2 Deviation and variance of $\hat{\boldsymbol{\gamma}}$ estimated in three situations ($n=100$)

情形	方法	$\hat{\gamma}_1/\times10^{-3}$		$\hat{\gamma}_2/\times10^{-3}$		$\hat{\gamma}_3/\times10^{-3}$		$\hat{\gamma}_4/\times10^{-3}$		$\hat{\gamma}_5/\times10^{-3}$	
		Bias	Var	Bias	Var	Bias	Var	Bias	Var	Bias	Var
情形(Ⅰ)	LSSVM	4.70	2.64	4.07	2.26	−3.91	2.34	−1.75	2.08	−5.07	2.35
	OLS	4.56	2.72	3.89	2.31	−4.31	2.37	−1.66	2.12	−5.02	2.39
情形(Ⅱ)	LSSVM	10.62	3.34	−1.60	2.79	−2.62	2.83	−0.38	3.29	3.18	3.03
	OLS	10.68	3.36	−2.27	2.65	−3.42	2.81	0.44	3.37	2.16	3.05
情形(Ⅲ)	LSSVM	2.46	2.48	−4.58	2.38	5.30	2.71	1.78	3.30	−2.13	2.61
	OLS	2.30	2.48	−4.53	2.42	5.64	2.74	2.65	3.35	−2.49	2.61

表 3　三种情形下 $\hat{\boldsymbol{\gamma}}$ 的偏差与方差($n=200$)

Tab. 3　Deviation and variance of $\hat{\boldsymbol{\gamma}}$ estimated in three situations ($n=200$)

情形	方法	$\hat{\gamma}_1/\times10^{-3}$		$\hat{\gamma}_2/\times10^{-3}$		$\hat{\gamma}_3/\times10^{-3}$		$\hat{\gamma}_4/\times10^{-3}$		$\hat{\gamma}_5/\times10^{-3}$	
		Bias	Var	Bias	Var	Bias	Var	Bias	Var	Bias	Var
情形(Ⅰ)	LSSVM	3.30	1.34	4.10	1.29	−2.92	1.13	−0.59	1.15	3.15	1.21
	OLS	3.38	1.35	4.07	1.29	−2.96	1.14	−0.52	1.16	3.09	1.22
情形(Ⅱ)	LSSVM	−5.15	1.00	−0.98	1.41	0.87	1.20	5.59	1.23	−0.94	1.08
	OLS	−4.84	0.98	−1.38	1.41	0.77	1.23	5.83	1.26	−0.90	1.09
情形(Ⅲ)	LSSVM	−3.65	1.19	−2.44	1.35	−3.42	1.43	−1.98	1.17	1.70	1.15
	OLS	−3.71	1.20	−2.51	1.36	−3.51	1.45	−1.91	1.18	1.65	1.16

表 4　回归系数函数 $\beta(t)$ 估计得到 $\hat{\boldsymbol{\gamma}}$ 的偏差与方差($n=500$)

Tab. 4　The deviation and variance of $\hat{\boldsymbol{\gamma}}$ estimated regression coofficiont fuction of $\beta(t)$ ($n=500$)

情形	方法	$\hat{\gamma}_1/\times10^{-3}$		$\hat{\gamma}_2/\times10^{-3}$		$\hat{\gamma}_3/\times10^{-3}$		$\hat{\gamma}_4/\times10^{-3}$		$\hat{\gamma}_5/\times10^{-3}$	
		Bias	Var	Bias	Var	Bias	Var	Bias	Var	Bias	Var
情形(Ⅰ)	LSSVM	−2.82	0.63	4.52	0.45	1.49	0.42	2.93	0.54	2.96	0.45
	OLS	−2.82	0.63	4.53	0.45	1.50	0.42	2.94	0.54	2.94	0.45
情形(Ⅱ)	LSSVM	0.98	0.45	−3.07	0.48	0.39	0.45	2.96	0.53	−1.00	0.44
	OLS	1.01	0.45	−3.11	0.48	0.44	0.45	2.99	0.53	−0.91	0.43
情形(Ⅲ)	LSSVM	−0.52	0.56	−1.64	0.42	0.12	0.53	−4.90	0.38	0.56	0.50
	OLS	−0.50	0.56	−1.67	0.42	0.13	0.53	−4.90	0.38	0.55	0.50

当 $n=100$ 时，情形(Ⅰ)～(Ⅲ)的某次模拟中 $\beta(t)$ 观测曲线及其估计曲线，如图 1～3 所示. 回归系数函数 $\beta(t)$ 估计得到的 $\hat{\beta}$ 平均偏离平方和的均值与方差，如表 5 所示.

由图 1 可知：当回归系数函数 $\beta(t)$ 恒为零时，因在函数型数据重构时选用傅里叶基，因此，两种方法的估计曲线在零附近但不能完全变为零. 结合表 5 情形(Ⅰ)中的结果可知：LSSVM 的 $\hat{\beta}$ 均值与方差比 OLS 要小.

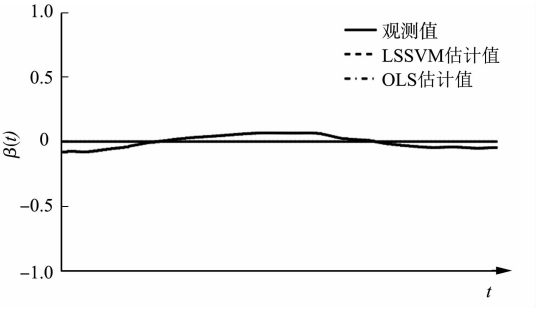

图 1　情形(Ⅰ)的某次模拟中 $\beta(t)$ 观测曲线及其估计曲线

Fig. 1　Observed and estimated curve of $\beta(t)$ in a simulation situation (Ⅰ)

由图 2 可知：当回归系数函数 $\beta(t)$ 设定为情形(Ⅱ)时，LSSVM 的估计曲线在峰值处偏离较小，更贴近观测曲线. 结合表 5 情形(Ⅱ)中的结果可知：$\hat{\beta}$ 平均偏离平方和的均值与方差随样本量的增加而递减，且 LSSVM 的表现均好于 OLS.

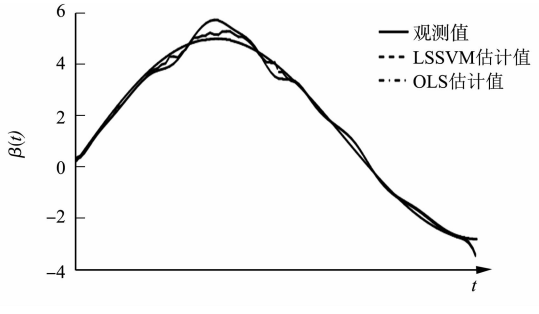

图 2　情形(Ⅱ)的某次模拟中 $\beta(t)$ 的观测曲线及其估计曲线

Fig. 2　Observed and estimated curves of $\beta(t)$ in simulation situation (Ⅱ)

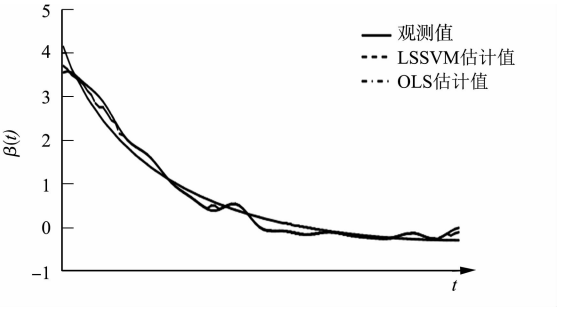

图 3　情形(Ⅲ)的某次模拟中 $\beta(t)$ 的观测曲线及其估计曲线

Fig. 3　Observed and estimated curves of $\beta(t)$ in simulation situation (Ⅲ)

由图 3 可知：当回归系数函数 $\beta(t)$ 设定为情形(Ⅲ)时，LSSVM 的估计曲线在头部和尾部更贴近观测曲线，其余两种方法估计相近都能较好地拟合观测曲线. 结合表 5 情形(Ⅲ)中的结果可知：当 $n=50$，

200时,LSSVM比OLS优势较大.

表5 $\hat{\beta}$的平均偏离平方和的均值与方差

Tab.5 Mean and variance of sum of mean deviation squares of $\hat{\beta}$

情形	方法	$n=50$		$n=80$		$n=200$		$n=500$	
		Mean($\hat{\beta}$)	Var($\hat{\beta}$)	Mean($\hat{\beta}$)	Var($\hat{\beta}$)	Mean($\hat{\beta}$)	Var($\hat{\beta}$)	Mean($\hat{\beta}$)	Var($\hat{\beta}$)
情形(Ⅰ)	LSSVM	1.725	1.155	1.188	0.594	0.740	0.269	0.355	0.067
	OLS	3.763	6.367	1.753	1.324	0.900	0.399	0.384	0.079
情形(Ⅱ)	LSSVM	4.078	6.500	2.735	4.279	1.427	1.774	1.124	1.420
	OLS	5.049	12.175	3.058	5.613	1.534	2.105	1.134	1.438
情形(Ⅲ)	LSSVM	2.099	1.915	1.329	0.774	0.891	0.383	0.413	0.121
	OLS	4.042	8.633	1.857	1.624	1.052	0.530	0.444	0.137

由表5可知:在不同样本量和不同回归系数函数设定下,LSSVM的$\hat{\beta}$均值与方差均小于OLS的值.综合系数向量$\boldsymbol{\gamma}$的估计结果看,在系数估计方面LSSVM与OLS相比更加准确稳健.

回归系数函数$\beta(t)$样本外预测值的RMSPE,如表6所示.由表6可知:在每个样本容量下,LSSVM的样本外预测误差比OLS小;在同一回归系数函数设定下,两种方法的预测误差随着样本量的增加略微上升,且LSSVM比OLS表现好.这说明LSSVM在系数估计上具有优势,有效提高了样本外预测的准确度.

表6 样本外预测值的RMSPE

Tab.6 RMSPE of out-of-sample predicted values

情形	方法	RMSPE			
		$n=50$	$n=100$	$n=200$	$n=500$
情形(Ⅰ)	LSSVM	0.434	0.468	0.480	0.493
	OLS	0.450	0.469	0.493	0.493
情形(Ⅱ)	LSSVM	0.437	0.476	0.492	0.494
	OLS	0.447	0.488	0.501	0.499
情形(Ⅲ)	LSSVM	0.431	0.466	0.484	0.502
	OLS	0.435	0.467	0.493	0.509

3 实例分析

3.1 数据预处理

以上证指数当日交易量和当日1 min高频交易价格数据作为次日上证指数开盘价的影响因素.由于每日的交易量数据过大,因此,将其取对数后作为离散型自变量$\boldsymbol{Z}_i$,当日1 min高频交易数据作为相依函数型自变量$X_i(t)$,次日的开盘价作为标量型应变量Y_{i+1},构建部分相依函数型线性模型,即

$$Y_{i+1}=\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i+\int_0^1\beta(t)X_i(t)\mathrm{d}t+\varepsilon_i,\qquad i=1,2,\cdots,n.$$

实例数据来源于锐思数据库,选取2018年1月至2018年12月的上证指数交易数据,包含次日开盘价、当日的交易量、及当日1 min高频交易数据.2018年共有243个交易日数据,删去最后1 d的交易日数据得到242个交易日数据,每个交易日有242个1 min高频交易价格数据.

3.2 上证指数开盘价预测

将前200个交易日数据作为训练样本,剩余42个交易日数据作为预测样本.分别使用文中提出的考虑函数型数据相依性的最小二乘支持向量机方法与未考虑相依性的最小二乘估计方法预测次日开盘价.预测结果与绝对误差的比较,如图4所示.

由图4可知:除个别交易日外,LSSVM估计的开盘价的绝对误差均OLS估计的开盘价的绝对误差,因此,文中方法的泛化能力更强.

为了综合比较预测效果,文中选取最大误差、最小误差、平均绝对误差、均方预测误差平方根评价方法的预测能力.两种方法预测结果的综合评价,如表7所示. 表7中:$E_{\max}$为最大误差;$E_{\min}$为最小误差;

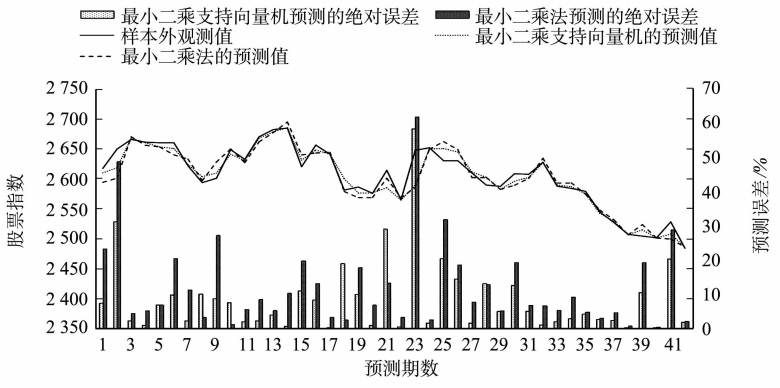

图 4　预测结果与绝对误差的比较

Fig. 4　Comparison of prediction results and absolute error

MAE 为平均绝对误差. 由表 7 可知：LSSVM 较好地预测次日的开盘价，其最大误差、最小误差、平均绝对误差、均方预测误差平方根均好于 OLS，由此证明 LSSVM 得到的预测效果优于 OLS 的预测效果.

表 7　两种方法预测结果的综合评价

Tab. 7　Comprehensive evaluation of prediction results of two methods

评价指标	LSSVM	OLS
$E_{\max}$	58.270	61.796
$E_{\min}$	0.209	0.341
MAE	8.438	12.175
RMSE	13.719	17.536

4　结束语

考虑到函数型数据的相依性结构特征，提出一种基于最小二乘支持向量机的部分相依函数型线性模型. 不同于其他的参数估计方法，利用无截断 Bartlett 核估计长期协方差函数，并将长期协方差函数所得到的特征函数对函数系数进行基展开，从而把函数系数的估计转化为参数向量的估计问题，随后运用最小二乘支持向量机给出了模型参数的估计. 通过数值模拟可知，与未考虑函数型数据相依性特征的最小二乘估计法相比，文中方法对向量系数的估计更加准确稳健，有效提高了样本外预测的准确度. 最后，将文中的参数估计方法应用于上证指数次日开盘价的预测中，进一步证明使用文中模型及参数估计方法的有效性和优越性.

参考文献：

[1] RAMSAY J O，SILVERMAN B W. Functionaldata analysis[M]. 2nd ed. New York：Springer-Verlag，2005. DOI：10.1002/0470013192. bsa239.

[2] HYNDMAN J，ULLAH S. Robust forecasting of mortality and fertility rates：A functional data approach[J]. Computational Statistics and Data Analysis，2007，51(10)：4942-4956. DOI：10.1016/j. csda. 2006. 07. 028.

[3] PARK J Y，QIAN Junhui. Functional regression of continuous state distributions[J]. Journal of Econometrics，2012，167(2)：397-412. DOI：10.1016/j. jeconom. 2011. 09. 024.

[4] 王国华. 中国股票市场日内波动率研究[D]. 武汉：中南财经政法大学，2017. DOI：10.1016/S0167-7152(99)00036-X.

[5] 苏桔芳，傅一铮，包浩华. 国债期限利差对经济周期转换的预警指示作用：基于函数型数据 Logit 模型的实证分析[J]. 宏观经济研究，2021(5)：20-30. DOI：10.16304/j. cnki. 11-3952/f. 2021. 05. 003.

[6] CARDOT H，FERRATY F，SARDA P. Functional linear model[J]. Statistics and Probability Letters，1999，45(1)：11-22. DOI：10.1016/S0167-7152(99)00036-X.

[7] CARDOT H，FERRATY F，SARDA P. Spline estimators for the functional linear model[J]. Statistica Sinica，2003，13(3)：571-591. DOI：10.1016/S0266-8920(03)00029-8.

[8] YAO Fang，MÜLLER H G，WANG J L. Functional data analysis for sparse longitudinal data[J]. Journal of the American Statistical Association，2005，100(470)：577-590. DOI：10.1198/016214504000001745.

[9] CRAMBES C，KNEIP A，SARDA P. Smoothing splines estimators for functional linear regression[J]. The Annals of Statistics，2009，37(1)：35-72. DOI：10. 1214/07-AOS563.

[10] HALL D P. Methodology and theory for partial least squares applied to functional data[J]. The Annals of Statis-

tics,2012,40:322-352. DOI: 10.2307/41713637.

[11] KALOGRIDIS I,AELST S V. Robust functional regression based on principal components[J]. Journal of Multivariate Analysis,2019,173:393-415. DOI:10.1016/j.jmva.2019.04.003.

[12] CRAMBES C,KNEIP A,SARDA P. Smoothing splines estimators for functional linear regression[J]. The Annals of Statistics,2009,37(1):35-72. DOI:10.1214/07-AOS563.

[13] ZHANG Daowen,LIN Xihong,SOWERS M F. Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome[J]. Biometrics,2007,63(2):351-362. DOI:10.1111/j.1541-0420.2006.00713.x.

[14] SHIN H. Partial functional linear regression[J]. Journal of Statistical Planning and Inference,2009,139(10):3405-3418. DOI:10.1016/j.jspi.2009.03.001.

[15] ZHOU Jianjun,CHEN Min. Spline estimators for semi-functional linear model[J]. Statistics and Probability Letters,2012,82(3):505-513. DOI:10.1016/j.spl.2011.11.027.

[16] 王晓光,高黎. 一类部分函数型线性模型的估计方法[J]. 高校应用数学学报:A辑,2013,28(3):253-265. DOI:10.3969/j.issn.1000-4424.2013.03.001.

[17] HÖRMANN S,KOKOSZKA P. Weakly dependent functional data[J]. The Annals of Statistics,2010,38(3):1845-1884. DOI:10.1214/09-aos768.

[18] KOKOSZKA P,REIMHERR M. Introduction to functional data analysis[M]. New York:CRC Press,2017.

[19] 李气芳. 相依函数型数据分析方法及其金融应用[D]. 厦门:华侨大学,2020. DOI:10.27155/d.cnki.ghqiu.2020.000517.

[20] KIEFER N M,VOGELSANG T J. Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation[J]. Econometrica,2002,70(5):2093-2095. DOI:10.1111/1468-0262.00366.

[21] 孟银凤,梁吉业. 基于最小二乘支持向量机的函数型数据回归分析[J]. 模式识别与人工智能,2014,27(12):1124-1130. DOI:10.3969/j.issn.1003-6059.2014.12.009.

[22] HÖRMANN S,KIDZISKI L,HALLIN M. Dynamic functional principal components[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology),2015,77(2):319-348. DOI:10.1111/rssb.12076.

[23] KOKOSZKA P,REIMHERR M. Introduction to functional data analysis[M]. Boca Raton:CRC Press,2017.

(**责任编辑**:陈志贤 **英文审校**:黄心中)