高校应用数学学报
2024, 39(4): 413-425

变系数部分函数型二元选择模型的统计推断

王龙兵[1], 张忠占[2]

(1. 衢州学院 教师教育学院, 浙江衢州 324000;

2. 北京工业大学 理学部, 北京 100124)

摘 要: 基于B样条sieve方法, 该文研究变系数部分函数型二元选择模型的估计及其渐近性质. 在一定的条件下, 证明了变系数函数的估计与斜率函数的估计的强相合性和渐近正态性. 在一定的条件下, 变系数函数的估计与斜率函数的估计达到最优收敛速度. 数值模拟和Tecator data的实证分析表明文中提出的估计方法是可行有效的.

关键词: 变系数回归模型; 函数型数据; 二元选择模型; B样条; 渐近性质

中图分类号: O212

文献标识码: A **文章编号**: 1000-4424(2024)04-0413-13

§1 引 言

随着科学技术的快速发展, 使得人们有能力收集到大量的密集数据(dense data), 这类数据也常常被称为曲线数据(curve data)或函数型数据(functional data).

近年来有不少函数型数据方面的研究成果, 如Hall等人[1]研究了函数型线性模型的最小二乘估计和岭估计及其收敛速度; Shin[2]推广了文献[1]中的模型, 提出了部分线性函数型模型, 并利用函数型主成分分析(functional principal components analysis)及最小二乘方法研究了回归参数和斜率函数(slope function)的估计并证明了回归参数的估计的渐近正态性, 得到了斜率函数的估计的收敛速度; Zhou等人[3]研究了文献[2]中的模型的推广模型, 即半函数线性模型(semi-functional linear model), 利用样条(spline)方法, 得到了非参函数(nonparametric function)与斜率函数估计; 鉴于分位数回归(quantile regression)的广泛应用, Lu等人[4]研究了函数部分线性分位数回归模型; 进一步, 文献[5-6]研究了函数型数据动态模型; 鉴于空间自回归模型(spatial autoregressive model)广泛的应用价值, 文献[7]推广了文献[2]中的模型, 研究了带有空间响应变量的部分函数型空间自回归模型, 更多的研究可以参考文献[8-11]等.

与线性回归模型相比较, 由Hastie等人[12]于1993年研究的变系数回归模型(varying coefficient regression model)能更加灵活地描述响应变量与协变量之间的关系, 因此受到大量的研究, 可以参考文献[13-17]等.

收稿日期: 2023-04-06 修回日期: 2024-10-07

基金项目: 国家自然科学基金(12271014); 衢州学院博士科研启动项目(BSYJ202117)

另一方面, 实际生活中有大量的需要做出二元选择的案例, 比如是否购入某种理财产品(比如股票、基金、国债、期货、期权、外汇)等, 该类问题就是属于计量经济学中著名的二元选择模型(binary choice model). 关于二元选择模型, 目前已有大量的研究, 可以参考文献[18-19]等.

综合以上分析, 本文研究变系数部分函数型二元选择模型

$$Y^* = Z_1^{\mathrm{T}}\alpha(Z_2) + \int_{\mathcal{T}} \beta(t)X(t)\mathrm{d}t + \epsilon, \quad Y = \begin{cases} 1, & \text{如果} Y^* > c; \\ 0, & \text{如果} Y^* \le c. \end{cases} \tag{1}$$

其中Y^*为潜变量(latent dependent variable), 不能直接观测到, 但能观测到Y, Z_1是p维实值随机向量, Z_2是定义在紧集$[e_1, e_2]$上的实值随机变量, 变系数$\alpha(\cdot) = (\alpha_1(\cdot), \cdots, \alpha_p(\cdot))^{\mathrm{T}}$是$p$维未知光滑函数向量, 斜率函数(slope function)$\beta(t)$是定义在区间$\mathcal{T}$上的平方可积函数, $X(t)$是一个定义在区间$\mathcal{T}$上的均值为0的平方可积的随机过程, c是决策临界点(critical decision point). ϵ为随机误差, 其分布函数为$F(\cdot)$, 若$\epsilon \sim$ Logis$(0,1)$, Logis$(0,1)$表示参数为(0, 1)的Logistic分布, 此时模型(1)表示变系数部分函数型Logistic模型; 若$\epsilon \sim N(0,1)$, 即标准正态分布, 此时模型(1)表示变系数部分函数型Probit模型, 当然随机误差也可以服从其他类型的分布, 比如$\epsilon \sim$ Cau$(0,1)$(参数为(0, 1)的Cauchy分布)等. 在数值模拟部分将对上述3种类型的随机误差分别进行模拟.

不失一般性, 设$\mathcal{T} = [0,1]$, 决策临界点$c = 0$. 本文研究目的是估计变系数$\alpha(\cdot)$和斜率函数$\beta(t)$, 给出估计的渐近性质, 并通过数值模拟研究该估计方法的有限样本性质.

§2 估计方法

由于变系数函数和斜率函数都是无穷维的, 不能直接估计, 需要转换为有限维空间进行估计, 鉴于B样条(B spline)的优良性质, 本文采用B样条来逼近变系数函数和斜率函数, 下面首先介绍一个B样条的重要引理.

引理1[20] 设$A = \{\beta(t) : \beta(t) \in C^m[0,1], \|\beta^{(j)}(t)\|_\infty \le I_j, j = 0, \cdots, m,$ $|\beta^{(m)}(t_1) - \beta^{(m)}(t_2)| \le I_{m+1}|t_1 - t_2|^\iota\}$, 其中$C^m[0,1]$为$[0,1]$区间上的$m$阶连续可导函数的全体, I_j是正的常数$(j = 0, \cdots, m+1)$, $0 < \iota \le 1$, $r = m + \iota$. 设$B(t) = (B_1(t), \cdots, B_N(t))^{\mathrm{T}}$为B样条基函数向量, k是节点数, l是样条的阶数, $N = k + l + 1$. 对于任意的$\beta(t) \in A$, 存在一个B样条逼近$b^{\mathrm{T}}B(t)$使得

$$\sup_{t\in[0,1]} \left|b^{\mathrm{T}}B(t) - \beta(t)\right| = O(k^{-r}),$$

这里$b = (b_1, \cdots, b_N)^{\mathrm{T}}$.

在模型(1)中, 设$\theta = (\alpha^{\mathrm{T}}(\cdot), \beta(\cdot))^{\mathrm{T}}$为模型参数, $\theta_0 = (\alpha_0^{\mathrm{T}}(\cdot), \beta_0(\cdot))^{\mathrm{T}}$为其真值. 设

$$\begin{aligned} \Theta_1 = \Big\{ & \alpha(\cdot) = (\alpha_1(\cdot), \cdots, \alpha_p(\cdot))^{\mathrm{T}} : \alpha_i(z) \in C^{m_1}[e_1, e_2], \left\|\alpha_i^{(j)}(z)\right\|_\infty \le I_j, \\ & i = 1, \cdots, p, j = 0, \cdots, m_1, \left|\alpha_i^{(m_1)}(z_1) - \alpha_i^{(m_1)}(z_2)\right| \le I_{m_1+1}\left|z_1 - z_2\right|^{\gamma_1} \Big\}, \end{aligned} \tag{2}$$

$$\begin{aligned} \Theta_2 = \Big\{ & \beta(t) : \beta(t) \in C^{m_2}[0,1], \left\|\beta^{(j)}(t)\right\|_\infty \le M_j, j = 0, \cdots, m_2, \\ & \left|\beta^{(m_2)}(t_1) - \beta^{(m_2)}(t_2)\right| \le M_{m_2+1}\left|t_1 - t_2\right|^{\gamma_2} \Big\}, \end{aligned} \tag{3}$$

上式中$\beta^{(j)}(t)$表示$\beta(t)$的第j阶导数, I_j和M_j是大于0的常数, $j = 0, \cdots, m_\rho + 1$. $0 < \gamma_\rho \le 1$, $\rho = 1, 2$. 因此模型(1)的参数空间可以记为

$$\Theta = \{\theta : \theta = (\alpha^{\mathrm{T}}(\cdot), \beta(\cdot))^{\mathrm{T}}, \alpha(\cdot) \in \Theta_1, \beta(\cdot) \in \Theta_2\} = \Theta_1 * \Theta_2.$$

因为参数空间Θ是无穷维的, 因此可以考虑sieve方法利用有限维参数空间Θ_n来逼近无穷维参数空间Θ, 然后在Θ_n上来估计θ, 即sieve方法的基本思想是将无穷维问题转化为有限维问题, 关于sieve方法的相关成果可参考文献[8, 21-22].

因为使用sieve方法来估计θ会涉及到逼近问题, 所以首先要定义一个合适的距离

$$d(\theta_1,\theta_2)=\left[\mathrm{E}(Z_1^{\mathrm{T}}(\alpha_1(Z_2)-\alpha_2(Z_2))+\langle\beta_1(t)-\beta_2(t),X(t)\rangle)^2\right]^{\frac{1}{2}}.$$

令

$$\Theta_{1n}=\Big\{\alpha_n(z_2)=(\alpha_{n1}(z_2),\cdots,\alpha_{np}(z_2)^{\mathrm{T}}:\alpha_{ni}(z_2)=\sum_{j=1}^{N_1}a_{ij}B_{1j}(z_2),\\ i=1,2,\cdots,p,\max_{j=1,\cdots,N_1}|a_{ij}|\le I\Big\},\tag{4}$$

$$\Theta_{2n}=\Big\{\beta_n(t):\beta_n(t)=\sum_{j=1}^{N_2}b_jB_{2j}(t),\max_{j=1,\cdots,N_2}|b_j|\le M\Big\},\tag{5}$$

(4)式中的I和(5)式中的M都是大于0的常数, $B_{1j}\in S_{m_1,k_1}, B_{2j}\in S_{m_2,k_2}, S_{m_\rho,k_\rho}$是阶数为$m_\rho$, 等距节点数为$k_\rho$的B-样条空间, $k_\rho=n^{\nu_\rho}(0<\nu_\rho<\frac{1}{2})$, $N_\rho=k_\rho+m_\rho+1$, $\rho=1,2$. k_1和k_2可以由AIC准则(Akaike information criterion)来确定.

设$\Theta_n=\Theta_{1n}*\Theta_{2n}$, $\pi_n\theta=(\alpha_n^{\mathrm{T}}(z_2),\beta_n(t))^{\mathrm{T}}$. 由引理1, 对任意$\theta\in\Theta$, 都存在$\pi_n\theta=(\alpha_n^{\mathrm{T}}(z_2),\beta_n(t))^{\mathrm{T}}\in\Theta_n$, 使得$d(\theta,\pi_n\theta)\le O(k_1^{-r_1}+k_2^{-r_2})$. 所以$\{\Theta_n\}$可以作为$\Theta$的sieve空间.

设$W_i=(Z_{1i}^{\mathrm{T}},Z_{2i},X_i(t),Y_i)^{\mathrm{T}}$是$W=(Z_1^{\mathrm{T}},Z_2,X(t),Y)^{\mathrm{T}}$的$n$个i.i.d(独立同分布)样本, $i=1,2,\cdots,n$, $\widetilde{W_n}=(W_1,\cdots,W_n)^{\mathrm{T}}$. 令

$$L_n(\theta;\widetilde{W_n})=P_nl(\theta;W)=\frac{1}{n}\sum_{i=1}^{n}l(\theta;W_i)=\\ \frac{1}{n}\sum_{i=1}^{n}\Big[Y_i\log F(Z_{1i}^{\mathrm{T}}\alpha(Z_{2i})+\langle\beta(t),X_i(t)\rangle)+\\ (1-Y_i)\log(1-F(Z_{1i}^{\mathrm{T}}\alpha(Z_{2i})+\langle\beta(t),X_i(t)\rangle))\Big]\tag{6}$$

为目标函数, 则

$$\hat{\theta}_n=(\hat{\alpha}_n^{\mathrm{T}}(\cdot),\hat{\beta}_n(\cdot))^{\mathrm{T}}=\arg\sup_{\theta\in\Theta_n}L_n(\theta;\widetilde{W_n})$$

是θ_0的sieve极大似然估计.

§3 渐近性质

为研究sieve极大似然估计的渐近性质, 本文做如下的假设.

C1: $\theta_0\in\Theta$.

C2: $\mathrm{E}\|X\|^2<C<\infty$, 其中$C>0$是常数.

C3: Z_2有紧支撑集$[e_1,e_2]$且Z_2的密度函数$f_{Z_2}(z_2)$满足

$$0<\inf_{z_2\in[e_1,e_2]}f_{Z_2}(z_2)\le\sup_{z_2\in[e_1,e_2]}f_{Z_2}(z_2)<\infty.$$

C4: 存在两个常数C_1和C_2使得$C_1\le Z_1^{\mathrm{T}}\alpha(Z_2)+\int_0^1\beta(t)X(t)\mathrm{d}t\le C_2$, a.s. P_{θ_0}.

C5: $r_1=r_2=r$, $k_1\sim k_2\sim k$, 其中$k=n^{\nu}(0<\nu<\frac{1}{2})$.

C6: $X(t)$的协方差算子的特征值严格大于0.

C7: 存在常数$\varepsilon>0$, 使得

$$\sup_{\alpha\in B(\alpha_0,\varepsilon),\ \beta\in B(\beta_0,\varepsilon)} P\left\{Z_1^{\mathrm{T}}\alpha(Z_2)\text{与}\int_0^1\beta(t)X(t)\mathrm{d}t\text{共线}\right\}<1,$$

其中$B(\beta_0,\varepsilon)$是以$\beta_0(t)$为中心的$L_\infty\varepsilon$-球.

注3.1 C1是回归模型的基本假设, 方便研究模型变系数函数和斜率函数估计的各类性质; C2是关于函数型数据$X(t)$的常见假设, 使得可以把斜率函数估计的各类性质建立在$L^2[0,1]$理论的基础上; C3是非参数回归模型的常见假设; C4在定理3.3中用到; C5是为了讨论的方便; C6和C7是为保证变系数函数和斜率函数的可识别性.

引理3.1[8] 设$\mathcal{F}_n=\{l(\theta,\cdot):\theta\in\Theta_n\}$, 则其覆盖数满足$N(\varepsilon,\mathcal{F}_n,L_\infty)\le G\left(\frac{1}{\varepsilon}\right)^{pN_1+N_2}$, 其中$G$是常数.

定理3.1 假设条件C1-C7成立, 则

$$\|\hat\alpha_n(\cdot)-\alpha_0(\cdot)\|_2\to 0,\qquad \|\hat\beta_n(\cdot)-\beta_0(\cdot)\|_2\to 0,$$

a.s. P_{θ_0}.

定理3.1的证明 该定理的证明类似于文献[26]中的定理3.1的证明.

定理3.2 假设条件C1-C7成立且节点数满足$k\sim n^{\frac{1}{1+2r}}$, 则

$$\|\hat\alpha_n(\cdot)-\alpha_0(\cdot)\|_2=O_p\left(n^{\frac{-r}{1+2r}}\right),\qquad \|\hat\beta_n(\cdot)-\beta_0(\cdot)\|_2=O_p\left(n^{\frac{-r}{1+2r}}\right).$$

定理3.2的证明 该定理的证明类似于文献[26]中的定理3.2的证明.

注3.2 由定理3.2可知变系数函数向量的估计$\hat\alpha_n(\cdot)$和斜率函数的估计$\hat\beta_n(\cdot)$达到Stone[23]提出的最优收敛速度(非参数回归样条估计的最优收敛速度).

下面研究变系数函数向量$\alpha(\cdot)$和斜率函数$\beta(t)$的估计的逐点渐近正态性, 为方便先给出以下记号.

$$B_1(z_2)=(B_{11}(z_2),\cdots,B_{1N_1}(z_2))^{\mathrm{T}},$$

$$\mathcal{B}_1(z_2)=\begin{pmatrix} B_1(z_2)^{\mathrm{T}} & 0 & \dots & 0\\ 0 & B_1(z_2)^{\mathrm{T}} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & B_1(z_2)^{\mathrm{T}}\end{pmatrix}_{p\times pN_1},$$

$$\Gamma(z_1,z_2)=(z_{11}B_1^{\mathrm{T}}(z_2),\cdots,z_{1p}B_1^{\mathrm{T}}(z_2))^{\mathrm{T}},\qquad B_2(t)=(B_{21}(t),\cdots,B_{2N_2}(t))^{\mathrm{T}},$$

$$\Upsilon(X)=\langle X,B_2(t)\rangle=(\langle X(t),B_{21}(t)\rangle,\cdots,\langle X(t),B_{2N_2}(t)\rangle)^{\mathrm{T}},$$

$$a_i=(a_{i1},\cdots,a_{iN_1})^{\mathrm{T}},a_i^*=(a_{i1}^*,\cdots,a_{iN_1}^*)^{\mathrm{T}},i=1,\cdots,p,$$

$$a=(a_1^{\mathrm{T}},\cdots,a_p^{\mathrm{T}})^{\mathrm{T}}_{pN_1\times1},b=(b_1,\cdots,b_{N_2})^{\mathrm{T}},$$

$$\alpha_0(z_2)=(\alpha_{01}(z_2),\cdots,\alpha_{0p}(z_2))^{\mathrm{T}},\qquad a^*=(a_1^{*\mathrm{T}},\cdots,a_p^{*\mathrm{T}})^{\mathrm{T}}_{pN_1\times1}.$$

假设a_i^*和$b^*=(b_1^*,\cdots,b_{N_2}^*)^{\mathrm{T}}$分别是$\alpha_{0i}(z_2)$和$\beta_0(t)$的$L_\infty$逼近B-样条系数向量, 满足

$$\sup_{z_2\in[e_1,e_2]}\left|a_i^{*\mathrm{T}}B_1(z_2)-\alpha_{0i}(z_2)\right|=O(k_1^{-r}),\qquad \sup_{t\in[0,1]}\left|b^{*\mathrm{T}}B_2(t)-\beta_0(t)\right|=O(k_2^{-r}),$$

$i=1,\cdots,p$.

由(6)可知, 系数函数向量$\alpha(\cdot)$和斜率函数$\beta(t)$的估计的系数向量可以由

$$\begin{aligned}(\hat a^{\mathrm{T}},\hat b^{\mathrm{T}})^{\mathrm{T}}=\arg\sup_{(a^{\mathrm{T}},b^{\mathrm{T}})^{\mathrm{T}}\in\mathbf{R}^{pN_1+N_2}}\frac{1}{n}\sum_{i=1}^n\Big\{&Y_i\log F(a^{\mathrm{T}}\Gamma(Z_{1i},Z_{2i})+b^{\mathrm{T}}\Upsilon(X_i))+\\&(1-Y_i)\log(1-F(a^{\mathrm{T}}\Gamma(Z_{1i},Z_{2i})+b^{\mathrm{T}}\Upsilon(X_i)))\Big\}\end{aligned}\tag{7}$$

得到.

定理3.3 假设条件C1-C7成立且节点数满足$\frac{n}{k^{1+2r}}=o(1)$, 则

$$\sqrt{\frac{n}{k_1}}(\hat{\alpha}_n(z_2)-\alpha^*(z_2))\xrightarrow{d}N(0,\Sigma_1(z_2)),\qquad \sqrt{\frac{n}{k_2}}(\hat{\beta}_n(t)-\beta^*(t))\xrightarrow{d}N(0,\Sigma_2(t)),$$

其中

$$\hat{\alpha}_n(z_2)=\begin{pmatrix}\hat{a}_1^{\mathrm{T}}B_1(z_2)\\ \vdots\\ \hat{a}_p^{\mathrm{T}}B_1(z_2)\end{pmatrix},\qquad \alpha^*(z_2)=\begin{pmatrix}a_1^{*\mathrm{T}}B_1(z_2)\\ \vdots\\ a_p^{*\mathrm{T}}B_1(z_2)\end{pmatrix}=\mathcal{B}_1(z_2)a^*,$$

$$\beta^*(t)=b^{*\mathrm{T}}B_2(t),\qquad \Sigma_1(z_2)=\lim_{n\to\infty}\frac{1}{k_1}\mathcal{B}_1(z_2)(H_1-H_3H_2^{-1}H_3^{\mathrm{T}})^{-1}\mathcal{B}_1(z_2)^{\mathrm{T}},$$

$$\Sigma_2(t)=\lim_{n\to\infty}\frac{1}{k_2}B_2(t)^{\mathrm{T}}(H_2-H_3^{\mathrm{T}}H_1^{-1}H_3)^{-1}B_2(t),$$

其他符号的含义见下面定理3.3的证明.

定理3.3的证明

设$\vartheta=(\vartheta_1^{\mathrm{T}},\vartheta_2^{\mathrm{T}})^{\mathrm{T}}\in\mathbf{R}^{pN_1+N_2},|\vartheta|_\infty<C$,

$$S_{1n}(\vartheta)=\sum_{i=1}^{n}\Big[l\big(\sqrt{\frac{k_1}{n}}\vartheta_1^{\mathrm{T}}\Gamma(Z_{1i},Z_{2i})+\sqrt{\frac{k_2}{n}}\vartheta_2^{\mathrm{T}}\Upsilon(X_i)+\omega_i\big)-l(\omega_i)\Big],\tag{8}$$

上式中

$$\omega_i=a^{*\mathrm{T}}\Gamma(Z_{1i},Z_{2i})+b^{*\mathrm{T}}\Upsilon(X_i),\qquad l(x)=Y\log F(x)+(1-Y)\log(1-F(x)),$$

不难验证$\hat{\vartheta}_1=\sqrt{\frac{n}{k_1}}\big(\hat{a}-a^*\big)$与$\hat{\vartheta}_2=\sqrt{\frac{n}{k_2}}\big(\hat{b}-b^*\big)$能使(8)式达到最大. 对函数$l(x)$进行Taylor展开可得

$$l(x+\Delta x)-l(x)=l'(x)\Delta x+\frac{l''(x)}{2}(\Delta x)^2+o((\Delta x)^2).\tag{9}$$

设

$$S_{2n}(\vartheta)=\vartheta_1^{\mathrm{T}}A_1+\vartheta_2^{\mathrm{T}}A_2+\frac{1}{2}\vartheta_1Q_1\vartheta_1+\frac{1}{2}\vartheta_2^{\mathrm{T}}Q_2\vartheta_2+\vartheta_1^{\mathrm{T}}Q_3\vartheta_2,\tag{10}$$

上式中

$$A_1=\frac{\sqrt{k_1}}{\sqrt{n}}\sum_{i=1}^{n}\Gamma(Z_{1i},Z_{2i})l'(\omega_i),\qquad A_2=\frac{\sqrt{k_2}}{\sqrt{n}}\sum_{i=1}^{n}\Upsilon(X_i)l'(\omega_i),$$

$$Q_1=\frac{k_1}{n}\sum_{i=1}^{n}\Gamma(Z_{1i},Z_{2i})\Gamma(Z_{1i},Z_{2i})^{\mathrm{T}}l''(\omega_i),\qquad Q_2=\frac{k_2}{n}\sum_{i=1}^{n}\Upsilon(X_i)\Upsilon(X_i)^{\mathrm{T}}l''(\omega_i),$$

$$Q_3=\frac{\sqrt{k_1k_2}}{n}\sum_{i=1}^{n}\Gamma(Z_{1i},Z_{2i})\Upsilon(X_i)^{\mathrm{T}}l''(\omega_i),$$

$$l'(\omega_i)=Y_i\frac{f(\omega_i)}{F(\omega_i)}+(1-Y_i)\frac{-f(\omega_i)}{1-F(\omega_i)},\tag{11}$$

$$l''(\omega_i)=Y_i\frac{f'(\omega_i)F(\omega_i)-f^2(\omega_i)}{F^2(\omega_i)}+(1-Y_i)\frac{-f'(\omega_i)(1-F(\omega_i))-f^2(\omega_i)}{(1-F(\omega_i))^2}.\tag{12}$$

综合条件$|\vartheta|_\infty<C$, 引理3.1和定理3.2可得

$$\sup_{\vartheta\in\Theta_n}|S_{1n}(\vartheta)-S_{2n}(\vartheta)|=o_p(1).\tag{13}$$

由(10)和(13)有

$$\begin{pmatrix}\hat{\vartheta}_1\\ \hat{\vartheta}_2\end{pmatrix}=\begin{pmatrix}Q_1 & Q_3\\ Q_3^{\mathrm{T}} & Q_2\end{pmatrix}^{-1}\begin{pmatrix}A_1\\ A_2\end{pmatrix}+o_p(1).\tag{14}$$

设

$$\omega_{0i} = Z_{1i}^{\mathrm{T}}\alpha_0(Z_{2i}) + \langle \beta_0, X_i\rangle,$$

由条件$\frac{n}{k^{1+2r}} = o(1)$有

$$\sup_{Z_{1i},Z_{2i}} |\omega_i - \omega_{0i}| \le O_p(k_1^{-r} + k_2^{-r}) = o_p(n^{\frac{-r}{1+2r}}),$$

由$l'(\cdot)$的连续性和定理的条件以及文献[24]的定理3, 可以忽略$l'(\omega_i)$与$l'(\omega_{0i})$之间的差异以及$l''(\omega_i)$与$l''(\omega_{0i})$之间的差异. 所以$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$的条件协方差可以利用

$$\Lambda = \begin{pmatrix} H_1 & H_3 \\ H_3^{\mathrm{T}} & H_2 \end{pmatrix} \tag{15}$$

表示, 上式中

$$H_1 = \frac{k_1}{n}\sum_{i=1}^{n} \Gamma(Z_{1i}, Z_{2i})\Gamma(Z_{1i}, Z_{2i})^{\mathrm{T}}\mathrm{E}l'^2(\omega_{0i}|Y_i, Z_{1i}, Z_{2i}, X_i(t)),$$

$$H_2 = \frac{k_2}{n}\sum_{i=1}^{n} \Upsilon(X_i)\Upsilon(X_i)^{\mathrm{T}}\mathrm{E}l'^2(\omega_{0i}|Y_i, Z_{1i}, Z_{2i}, X_i(t)),$$

$$H_3 = \frac{\sqrt{k_1k_2}}{n}\sum_{i=1}^{n} \Gamma(Z_{1i}, Z_{2i})\Upsilon(X_i)^{\mathrm{T}}\mathrm{E}l'^2(\omega_{0i}|Y_i, Z_{1i}, Z_{2i}, X_i(t)).$$

简单计算可知

$$\mathrm{E}l'^2(\omega_{0i}|Y_i, Z_{1i}, Z_{2i}, X_i(t)) = -\mathrm{E}l''(\omega_{0i}|Y_i, Z_{1i}, Z_{2i}, X_i(t)) = \frac{f^2(\omega_{0i})}{F(\omega_{0i})(1-F(\omega_{0i}))},$$

因此

$$\begin{pmatrix} \mathrm{E}(Q_1|Y_i, Z_{1i}, Z_{2i}, X_i(t)) & \mathrm{E}(Q_3|Y_i, Z_{1i}, Z_{2i}, X_i(t)) \\ \mathrm{E}(Q_3^{\mathrm{T}}|Y_i, Z_{1i}, Z_{2i}, X_i(t)) & \mathrm{E}(Q_2|Y_i, Z_{1i}, Z_{2i}, X_i(t)) \end{pmatrix} = -\Lambda + o_p(1). \tag{16}$$

注意到

$$\sqrt{\frac{n}{k_1}}(\hat{\alpha}_n(z_2) - \alpha^*(z_2)) = \sqrt{\frac{n}{k_1}}\mathcal{B}_1(z_2)(\hat{a} - a^*), \qquad \sqrt{\frac{n}{k_2}}(\hat{\beta}_n(t) - \beta^*(t)) = \sqrt{\frac{n}{k_2}}(\hat{b} - b^*)^{\mathrm{T}}B_2(t),$$

因此对于任意的$z_2 \in [e_1, e_2], t \in [0,1]$, 当$n \to \infty$有

$$\sqrt{\frac{n}{k_1}}(\hat{\alpha}_n(z_2) - \alpha^*(z_2)) \xrightarrow{d} N(0, \Sigma_1(z_2)),$$

$$\sqrt{\frac{n}{k_2}}(\hat{\beta}_n(t) - \beta^*(t)) \xrightarrow{d} N(0, \Sigma_2(t)),$$

上式中

$$\Sigma_1(z_2) = \lim_{n\to\infty}\frac{1}{k_1}\mathcal{B}_1(z_2)(H_1 - H_3H_2^{-1}H_3^{\mathrm{T}})^{-1}\mathcal{B}_1(z_2)^{\mathrm{T}},$$

$$\Sigma_2(t) = \lim_{n\to\infty}\frac{1}{k_2}B_2(t)^{\mathrm{T}}(H_2 - H_3^{\mathrm{T}}H_1^{-1}H_3)^{-1}B_2(t),$$

§4 数值模拟

下面利用数值模拟来研究所提出的估计的有限样本性质, 计算和画图均使用R软件, 模拟设计为

$$Y^* = Z_1\alpha_1(Z) + Z_2\alpha_2(Z) + \int_0^1 \beta_0(t)X(t)\mathrm{d}t + \epsilon,$$

$$Y = \begin{cases} 1, & \text{如果} Y^* > 0; \\ 0, & \text{如果} Y^* \le 0. \end{cases}$$

其中$Z_1 \sim N(0,1), Z_2 \sim Bi(1,0.5), Z \sim U[0,1], \alpha_1(z) = \sin(2\pi z), \alpha_2(z) = \cos(2\pi z)$, $\beta_0(t) = \sqrt{2}\sin(\frac{1}{2}\pi t) + 3\sqrt{2}\sin(\frac{3}{2}\pi t)$, $X(t) = \sum_{j=1}^{100} U_j\phi_j(t)$, 这里$Bi(1,0.5)$表示两点分布，其中取1的概率为0.5, U_j是相互独立的正态随机变量, 均值为0, 方差为$\lambda_j = \frac{25}{2}\big[(j-0.5)\pi\big]^{-2}$, $\phi_j(t) = \sqrt{2}\sin((j-0.5)\pi t)$, ϵ为随机误差.

为更好地说明估计方法, 在模拟中考虑以下中情形的随机误差. (1) $\epsilon \sim \text{Logis}(0,1)$, 此时模型(1)表示变系数函数型Logistic模型; (2) $\epsilon \sim N(0,1)$, 此时模型(1)表示变系数函数型Probit模型; (3) $\epsilon \sim \text{Cau}(0,1)$.

模拟中利用均匀节点数为4(即$k=2$)的三次B-样条来逼近$\alpha_1(z)$, $\alpha_2(z)$以及$\beta(t)$. 样本量为$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$, 分别重复1000次. 计算

$$\text{MSE}(\hat{\beta}(t)) = \frac{\sum_{i=1}^{n_{\text{grid1}}}(\hat{\beta}(t_i) - \beta(t_i))^2}{n_{\text{grid1}}}, \qquad \text{MSE}(\hat{\alpha}_1(z)) = \frac{\sum_{q=1}^{n_{\text{grid2}}}(\hat{\alpha}_1(z_q) - \alpha_1(z_q))^2}{n_{\text{grid2}}},$$

以及

$$\text{MSE}(\hat{\alpha}_2(z)) = \frac{\sum_{q=1}^{n_{\text{grid2}}}(\hat{\alpha}_2(z_q) - \alpha_2(z_q))^2}{n_{\text{grid2}}}$$

来研究估计的有限样本性质, 其中$n_{\text{grid1}} = n_{\text{grid2}} = 100$, $\{t_i, i = 1,2,\cdots,n_{\text{grid1}}\}$和$\{z_q, q = 1,2,\cdots,n_{\text{grid2}}\}$是$[0,1]$区间上的均匀格子点. 不同随机误差不同样本量下的模拟结果在表1-表3中, $\beta(t)$, $\alpha_1(z)$以及$\alpha_2(z)$的估计曲线见图1-图9.

由表1-表3可知, 随着样本量的增大, 三个MSEs的均值和中位数以及标准差都变小, 结合图1-图9, 可知所提出的估计有较好的有限样本性质.

对比表1与表4, 表2与表5的模拟结果, 可以发现本文B样条方法比函数型主成分方法效果好.

表1 $\epsilon \sim \text{Logis}(0,1)$的模拟结果

样本量	$n_1 = 300$	$n_2 = 500$	$n_3 = 1000$
MSE($\hat{\beta}(t)$)的均值	5.6720	3.2172	1.3020
MSE($\hat{\beta}(t)$)的中位数	3.7030	2.0141	0.9218
MSE($\hat{\beta}(t)$)的标准差	6.2983	3.6597	1.2572
MSE($\hat{\alpha}_1(z)$)的均值	0.3561	0.1726	0.0737
MSE($\hat{\alpha}_1(z)$)的中位数	0.2852	0.1472	0.0629
MSE($\hat{\alpha}_1(z)$)的标准差	0.2837	0.1205	0.0486
MSE($\hat{\alpha}_2(z)$)的均值	0.7144	0.3182	0.1363
MSE($\hat{\alpha}_2(z)$)的中位数	0.5331	0.2649	0.1212
MSE($\hat{\alpha}_2(z)$)的标准差	1.0162	0.2332	0.0836

表2 $\epsilon \sim N(0,1)$的模拟结果

样本量	$n_1 = 300$	$n_2 = 500$	$n_3 = 1000$
MSE($\hat{\beta}(t)$)的均值	4.5389	2.0405	0.7682
MSE($\hat{\beta}(t)$)的中位数	2.8790	1.3278	0.5239
MSE($\hat{\beta}(t)$)的标准差	5.4739	2.3451	0.7317
MSE($\hat{\alpha}_1(z)$)的均值	0.3071	0.1140	0.0428
MSE($\hat{\alpha}_1(z)$)的中位数	0.2030	0.0836	0.0368
MSE($\hat{\alpha}_1(z)$)的标准差	0.4239	0.1017	0.0275
MSE($\hat{\alpha}_2(z)$)的均值	0.5088	0.2041	0.0794
MSE($\hat{\alpha}_2(z)$)的中位数	0.3707	0.1608	0.0682
MSE($\hat{\alpha}_2(z)$)的标准差	0.5268	0.2318	0.0511

表3 $\epsilon \sim \mathrm{Cau}(0,1)$的模拟结果

样本量	$n_1=300$	$n_2=500$	$n_3=1000$
MSE($\hat{\beta}(t)$)的均值	13.4018	5.3993	1.8997
MSE($\hat{\beta}(t)$)的中位数	8.0855	3.4611	1.3086
MSE($\hat{\beta}(t)$)的标准差	18.5722	6.1330	1.8009
MSE($\hat{\alpha}_1(z)$)的均值	0.9362	0.3306	0.1125
MSE($\hat{\alpha}_1(z)$)的中位数	0.5955	0.2464	0.0944
MSE($\hat{\alpha}_1(z)$)的标准差	1.3019	0.3208	0.0772
MSE($\hat{\alpha}_2(z)$)的均值	1.7856	0.5827	0.1988
MSE($\hat{\alpha}_2(z)$)的中位数	1.0717	0.4258	0.1610
MSE($\hat{\alpha}_2(z)$)的标准差	3.7512	0.5866	0.1395

表4 $\epsilon \sim \mathrm{Logis}(0,1)$, 函数型主成分方法的模拟结果

样本量	$n_1=300$	$n_2=500$	$n_3=1000$
MSE($\hat{\beta}(t)$)的均值	13.0916	6.5738	2.7447
MSE($\hat{\beta}(t)$)的中位数	11.0407	5.5547	2.3731
MSE($\hat{\beta}(t)$)的标准差	8.8430	4.2926	1.6917
MSE($\hat{\alpha}_1(z)$)的均值	0.3958	0.1820	0.0756
MSE($\hat{\alpha}_1(z)$)的中位数	0.3103	0.1545	0.0649
MSE($\hat{\alpha}_1(z)$)的标准差	0.3199	0.1280	0.0501
MSE($\hat{\alpha}_2(z)$)的均值	0.7785	0.3358	0.1395
MSE($\hat{\alpha}_2(z)$)的中位数	0.5644	0.2791	0.1236
MSE($\hat{\alpha}_2(z)$)的标准差	1.0611	0.2492	0.0858

表5 $\epsilon \sim N(0,1)$, 函数型主成分方法的模拟结果

样本量	$n_1=300$	$n_2=500$	$n_3=1000$
MSE($\hat{\beta}(t)$)的均值	10.6443	4.3137	1.6190
MSE($\hat{\beta}(t)$)的中位数	7.6831	3.4563	1.4340
MSE($\hat{\beta}(t)$)的标准差	11.4222	3.1327	1.0071
MSE($\hat{\alpha}_1(z)$)的均值	0.3967	0.1295	0.0450
MSE($\hat{\alpha}_1(z)$)的中位数	0.2412	0.0944	0.0380
MSE($\hat{\alpha}_1(z)$)的标准差	0.6097	0.1193	0.0288
MSE($\hat{\alpha}_2(z)$)的均值	0.6384	0.2265	0.0828
MSE($\hat{\alpha}_2(z)$)的中位数	0.4458	0.1775	0.0710
MSE($\hat{\alpha}_2(z)$)的标准差	0.7203	0.2554	0.0541

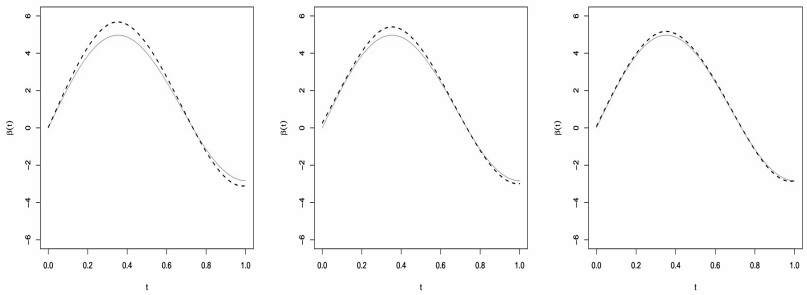

图1 $\epsilon \sim \mathrm{Logis}(0,1)$时$\beta_0(t)$与估计曲线(虚线), 从左到右分别对应于$n_1=300$, $n_2=500$, $n_3=1000$

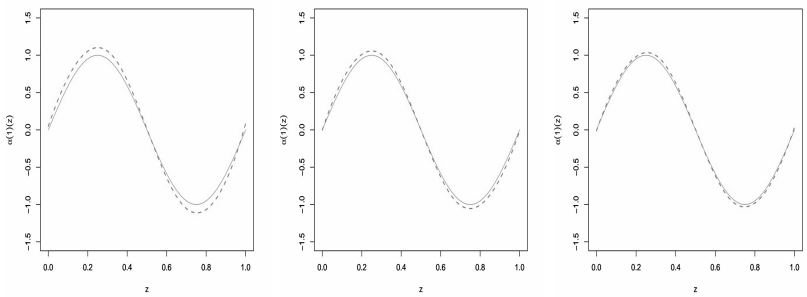

图2 $\epsilon \sim \text{Logis}(0,1)$时$\alpha_1(z)$与估计曲线(虚线), 从左到右分别对应于$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$

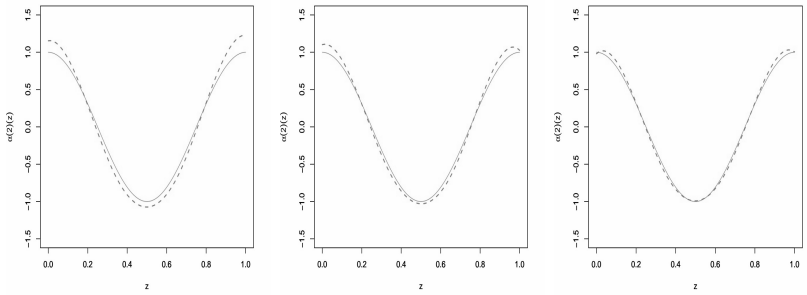

图3 $\epsilon \sim \text{Logis}(0,1)$时$\alpha_2(z)$与估计曲线(虚线), 从左到右分别对应于$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$

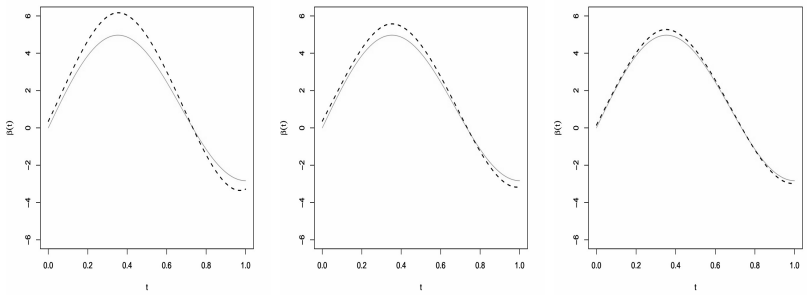

图4 $\epsilon \sim N(0,1)$时$\beta_0(t)$与估计曲线(虚线), 从左到右分别对应于$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$

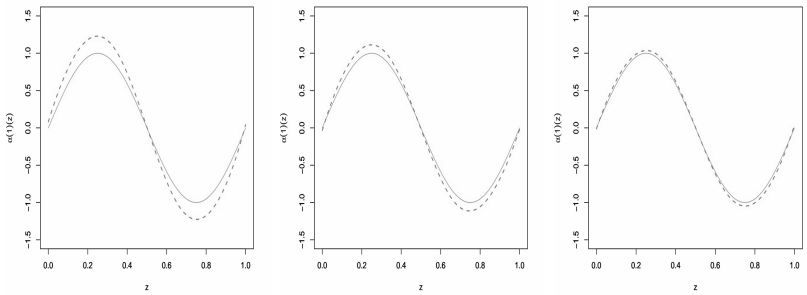

图5 $\epsilon \sim N(0,1)$时$\alpha_1(z)$与估计曲线(虚线), 从左到右分别对应于$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$

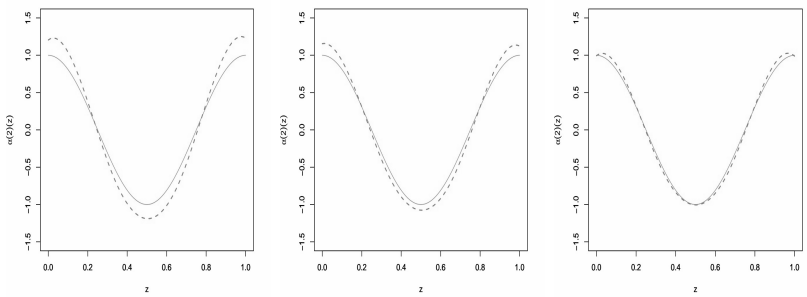

图6 $\epsilon \sim N(0,1)$时$\alpha_2(z)$与估计曲线(虚线), 从左到右分别对应于$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$

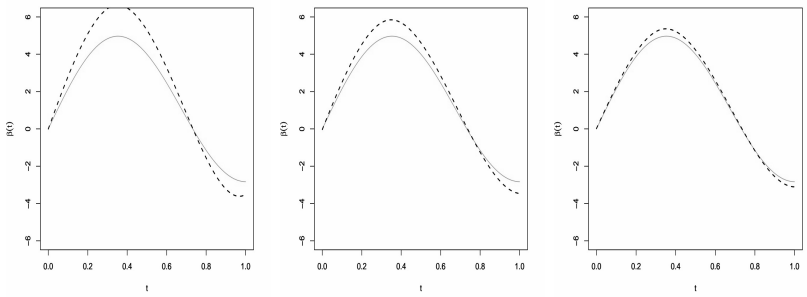

图7 $\epsilon \sim \mathrm{Cau}(0,1)$时$\beta_0(t)$与估计曲线(虚线), 从左到右分别对应于$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$

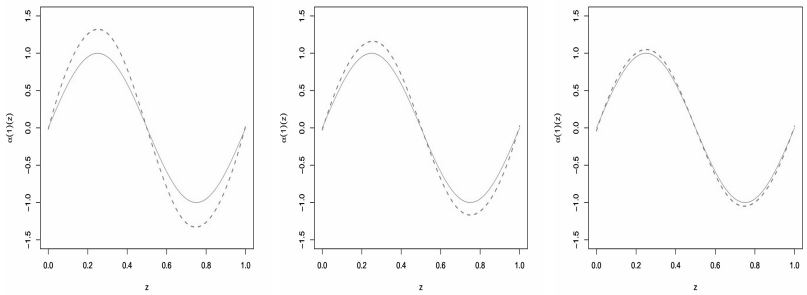

图8 $\epsilon \sim \mathrm{Cau}(0,1)$时$\alpha_1(z)$与估计曲线(虚线), 从左到右分别对应于$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$

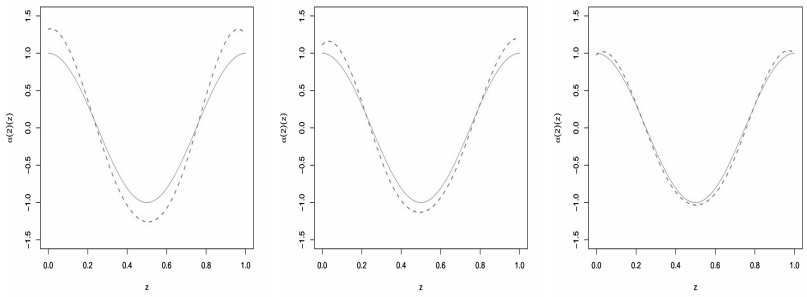

图9 $\epsilon \sim \mathrm{Cau}(0,1)$时$\alpha_2(z)$与估计曲线(虚线), 从左到右分别对应于$n_1 = 300$, $n_2 = 500$, $n_3 = 1000$

§5 实际数据分析

Tecator data(http: //lib.stat.cmu.edu/datasets/tecator)受到广泛研究, 如文献[8, 25]. 该数据集包含215个肉质品观测样本数据, 每个样本数据包括波段850-1050纳米范围内的100个频道光谱以及蛋白质含量(content of protein)、水分含量(content of water)、脂肪含量(content of fat). 本节利用本文的模型与方法研究Tecator data. 计算和画图均使用R软件, 模型

$$Y^* = \alpha(Z_1)Z_2 + \int_{850}^{1050} \beta_0(t)X(t)\mathrm{d}t + \epsilon, \qquad Y = \begin{cases} 1, & \text{如果} Y^* > 18.5; \\ 0, & \text{如果} Y^* \le 18.5. \end{cases}$$

Y^*表示蛋白质含量, Z_1表示水分含量, Z_2表示脂肪含量, $X(t)$是函数型变量(functional variable), 表示光谱曲线, $\epsilon \sim \mathrm{Logis}(0,1)$. $\alpha(Z_1)$, $\beta_0(t)$是2个待估的未知函数.

估计曲线见下图, 从$\alpha(Z_1)$的估计曲线看出, 具有明显的非线性特征, 若用线性模型则难以捕捉到这个信息, 表明本文的变系数模型对数据集Tecator data的建模是合适的.

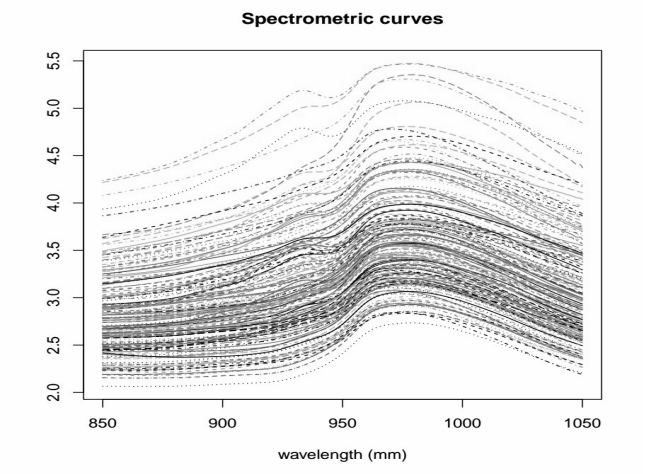

图10 光谱曲线

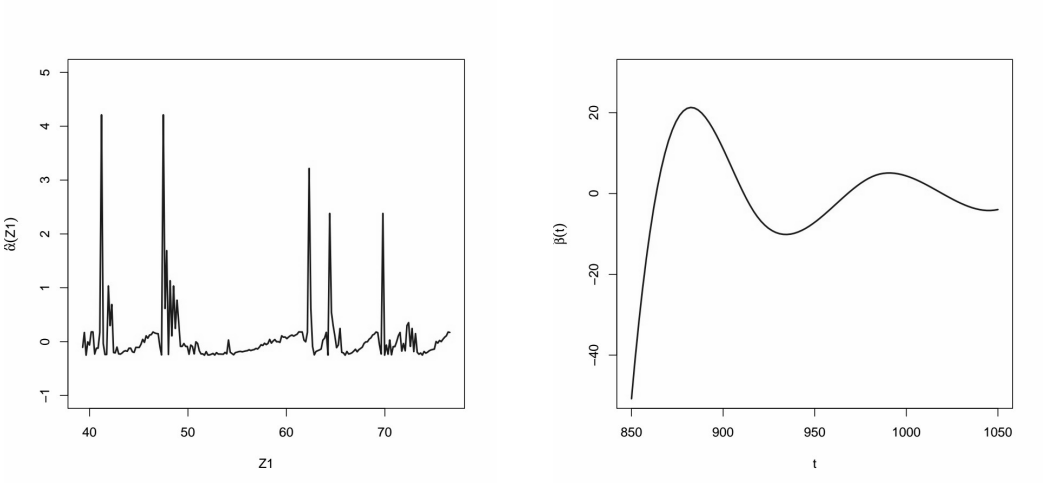

图11 $\alpha(Z_1)$的估计曲线(左) 与$\beta_0(t)$的估计曲线(右)

§6 结论

本文研究变系数部分函数型二元选择模型的估计. 在一定的条件下, 证明了该估计的强相合性和变系数函数与斜率函数的逐点渐近正态性, 得到了变系数函数与斜率函数的最优收敛速度. 通过数值模拟可知, 在3种不同的随机误差下, 本文的估计效果都较好, 这为后续的实际应用提供了理论依据.

参考文献:

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

[2] Shin H. Partial functional linear regression[J]. Journal of Statistical Planning and Inference 2009, 139(10): 3405-3418.

[3] Zhou Jianjun, Chen Min. Spline estimators for semi-functional linear model[J]. Statistics & Probability Letters, 2012, 82(3): 505-513.

[4] Lu Ying, Du Jiang, Sun Zhimeng. Functional partially linear quantile regression model[J]. Metrika, 2014, 77(2): 317-332.

[5] Ma Haiqiang, Bai Yang, Zhu Zhongyi. Dynamic single-index model for functional data[J]. Science China: Mathematics, 2016, 59(12): 2561-2584.

[6] Ma Haiqiang, Zhu Zhongyi. Continuously dynamic additive models for functional data[J]. Journal of Multivariate Analysis, 2016, 150: 1-13.

[7] 徐登可, 田瑞琴. 函数型空间自回归模型的贝叶斯估计[J]. 高校应用数学学报, 2022, 37(3): 323-336.

[8] Huang Lele, Wang Huiwen, Cui Hengjian, et al. Sieve M-estimator for a semi-functional linear model[J]. Science China: Mathematics, 2015, 58(11): 2421-2434.

[9] Fan Jianqing , Zhang Jinting. Two-step estimation of functional linear models with applications to longitudinal data[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2000, 62(2): 303-322.

[10] Cardot H, Ferraty F, Sarda P. Spline estimators for the functional linear model[J]. Statistica Sinica, 2007, 13(3): 571-591.

[11] Ramsay J O, Silverman B W. Functional Data Analysis[M]. New York: Springer, 1997.

[12] Hastie T, Tibshirani R. Varying-coefficient models[J]. Journal of the Royal Statistical Society, 1993, 55(4): 757-796.

[13] 唐庆国, 晋鹏. 空间半参数变系数部分线性分位数回归中的B-样条估计法[J]. 统计与信息论坛, 2018, 33(6): 9-13.

[14] lv Xiaoling, Wang Xiaoning, Sun Zhimeng. FIC based model averaging for the censored quantile varying coefficient regression model[J]. Journal of Systems Science and Mathematical Sciences, 2018, 38(7): 746-763.

[15] 赵静, 蒲越. 空间滞后单指标变系数模型的估计及其应用[J]. 数量经济技术经济研究, 2021, 38(11): 163-181.

[16] 孙怡帆, 王彩晶, 罗梓烨. 基于变系数模型的高维数据异同性识别方法研究[J]. 统计研究, 2021, 38(5): 136-146.

[17] 薛留根. 现代统计模型[M]. 北京: 科学出版社, 2012.

[18] 邸俊鹏, 张晓峒. 二元选择分位数回归模型的贝叶斯估计方法及模拟研究[J]. 统计与决策, 2019, 35(5): 11-16.
[19] 纪园园, 王黎明, 张杭辉, 等. 非参数选择机制下二值因变量的内生选择模型的估计[J]. 中国科学: 数学, 2022, 52(3): 307-330.
[20] Schumaker L L. Spline Functions: Basic Theory[M]. New York: Wiley, 1981.
[21] Xue Hongqi, Lam K F, Li Guoying. Sieve maximum likelihood estimator for semiparametric regression models with current status data[J]. Journal of the American Statistical Association, 2004, 99(466): 346-357.
[22] Shen Xiaotong, Wong Winghung. Convergence rate of sieve estimates[J]. Annals of Statistics, 1994, 22(2): 580-615.
[23] Stone C J. Optimal rates of convergence for nonparametric estimators[J]. Annals of Statistics, 1982, 10(4): 1040-1053.
[24] Yu Ping, Du Jiang, Zhang Zhongzhan. Single-index partially functional linear regression model[J]. Statistical Papers, 2020, 61: 1107-1123.
[25] Yu Ping, Du Jiang, Zhang Zhongzhan. Varying-coefficient partially functional linear quantile regression models[J]. Journal of the Korean Statistical Society, 2017, 46(3): 462-475.
[26] 王龙兵, 张忠占. 基于sieve方法的响应变量为当前状态数据的部分函数型线性模型的估计[J]. 高校应用数学学报, 2019, 34(1): 1-10.

Statistical inference for partial functional varying coefficient binary choice model

WANG Long-bing[1], ZHANG Zhong-zhan[2]
(1. College of Teacher Education, Quzhou University, Quzhou 324000, China;
2. Faculty of Science, Beijing University of Technology, Beijing 100124, China)

Abstract: In this paper, based on the Bspline sieve method, the estimation and asymptotic properties of the partial functional varying coefficient binary choice model are studied. Under certain conditions, the strong consistency and asymptotic normality of the estimates of varying coefficient functions and slope function are proved. Under certain conditions, the estimation of the varying coefficient function and the estimation of the slope function achieve the optimal convergence rate. Numerical simulation and empirical analysis of Tecator data show that the proposed estimation method is feasible and effective.

Keywords: varying coefficient regression model; functional data; binary choice model; B spline; asymptotic properties

MR Subject Classification: 62G08; 62G20