DOI :10.13203/j.whugis2004.07.022
第29卷 第7期 武汉大学学报·信息科学版 Vol. 29 No. 7
2004年7月 Geomatics and Information Science of Wuhan University July 2004

文章编号: 1671-8860(2004)07-0654-04 文献标识码: A

半参数回归模型中非参数信号的推估
——三次样条函数插值法

胡宏昌[1,2] 孙海燕[1]

(1 武汉大学测绘学院, 武汉市珞喻路 129 号, 430079)
(2 湖北师范学院数学系, 黄石市沈家营, 435002)

摘 要: 对于半参数回归模型 $y_i=\boldsymbol{X}_i^{\mathrm{T}}\beta+s(t_i)+e_i$, 在文献[1]的基础上考虑其非参数信号 $s(t)$ 的推估 $\hat{s}(t)$, 从而有利于研究非参数信号; 接着给出了 $s(t)$ 与 $\hat{s}(t)$ 的偏差大小, 从而为推估的准确性提供理论依据; 并以实例加以证明。

关键词: 半参数回归模型; 三次样条函数; 推估
中图法分类号: P207

文献[1]讨论了半参数模型:

$$y_i=\boldsymbol{X}_i^{\mathrm{T}}\beta+s(t_i)+e_i,\ i=1,2,\cdots,n$$

式中, y_i 为观测值; $\boldsymbol{X}_i$ 为 $p\times 1$ 维设计向量; $\beta=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ 为 $p\times 1$ 维待估参数向量; $n>p$, $e_i\sim N(0,\sigma_i^2)$ 且相互独立; $s\in R^1$ 为未知且满足一定条件的函数信号。

利用补偿最小二乘法则 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{V}+\alpha\boldsymbol{S}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{S}=\min$, 可得到参数 β 及非参数 $s(t_i)$ 的估计。该文的结果是非常有效的, 但对于非参数信号只给出了 $\hat{s}(t_i)$, 而没有得到 $s(t)(t\in[t_1,t_n])$ 的推估表达式 $\hat{s}(t)$, 这不利于人们研究非参数信号, 进而影响 y 的推估和预测。

1 非参数信号的推估

若能求出 $\hat{s}'(t)$ 在各个节点处的 $m_i=\hat{s}'(t_i)$, 则易得 $\hat{s}(t)$。实际上, 当 $m_i=\hat{s}'(t_i)$, $\hat{s}(t)$ 在子区间 $[t_{i-1},t_i]$ 上就是一个满足条件

$$\begin{aligned}&\hat{s}(t_{i-1})=u_{i-1},\ \hat{s}(t_i)=u_i,\\&\hat{s}'(t_{i-1})=m_{i-1},\ \hat{s}'(t_i)=m_i\end{aligned}\tag{1}$$

的三次 Hermite 插值多项式。由两点三次 Hermite 插值公式[2]可得:

$$\begin{aligned}\hat{s}(t)=&m_{i-1}\frac{(t_i-t)^2(t-t_{i-1})}{h_i^2}-\\&m_i\frac{(t-t_{i-1})^2(t_i-t)}{h_i^2}+\\&u_{i-1}\frac{(t_i-t)^2(2(t-t_{i-1})+h_i)}{h_i^3}+\\&u_i\frac{(t-t_{i-1})^2(2(t_i-t)+h_i)}{h_i^3},\\&t\in[t_{i-1},t_i],\ i=1,2,\cdots,n\end{aligned}\tag{2}$$

式中, $h_i=t_i-t_{i-1}$。

为了求得 m, 要利用 $\hat{s}(t)$ 的二阶导数在内节点 $t_i(i=1,2,\cdots,n-1)$ 上连续的条件。由式(2)可得:

$$\begin{aligned}\hat{s}''(t)=&-2m_{i-1}\cdot\frac{2t_i+t_{i-1}-3t}{h_i^2}-\\&2m_i\cdot\frac{2t_{i-1}+t_i-3t}{h_i^2}+6\cdot\frac{u_i-u_{i-1}}{h_i^3}\cdot\\&(t_i+t_{i-1}-2t)\end{aligned}\tag{3}$$

由式(3)得:

$$\hat{s}''(t_i-0)=\frac{2m_{i-1}}{h_i}+\frac{4m_i}{h_i}-6\frac{u_i-u_{i-1}}{h_i^2}$$

$$\hat{s}''(t_i+0)=-\frac{4m_i}{h_{i+1}}-\frac{2m_{i+1}}{h_{i+1}}+6\frac{u_{i+1}-u_i}{h_{i+1}^2}$$

由于 $\hat{s}''$ 在 t_i 上连续, 于是, 由 $\hat{s}''(t_i-0)=\hat{s}''(t_i+0)$ 可得:

$$\lambda_i m_{i-1}+2m_i+\mu_i m_{i+1}=c_i(i=1,2,\cdots,n-1)\tag{4}$$

收稿日期: 2004-04-26。
项目来源: 国家自然科学基金资助项目(40274005); 湖北省教育厅重大研究资助项目(2001Z06003)。

式中, $\lambda_i=\frac{h_{i+1}}{h_i+h_{i+1}}$; $c_i=3\left(\lambda_i\frac{u_i-u_{i-1}}{h_i}+\mu_i\frac{u_{i+1}-u_i}{h_{i+1}}\right)$; $\mu_i=\frac{h_i}{h_i+h_{i+1}}$。

为了解出方程组(4)的惟一解, 需用到边界条件:

$$s''(t_0)=u_0'',\ s''(t_n)=u_n'' \tag{5}$$

由式(3)、式(5)可得:

$$2m_0+m_1=3\frac{u_1-u_0}{h_1}-\frac{h_1}{2}\cdot u_0''\triangleq c_0 \tag{6}$$

$$m_{n-1}+2m_n=3\frac{u_n-u_{n-1}}{h_n}-\frac{h_n}{2}\cdot u_n''\triangleq c_n \tag{7}$$

由式(4)~式(7)可确定 $m_i(i=0, 1, 2, \cdots, n)$的线性方程组:

$$\begin{pmatrix} 2 & 1 & & & \\ \lambda_1 & 2 & \mu_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda_{n-1} & 2 & \mu_{n-1} \\ & & & 1 & 2 \end{pmatrix}\begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ m_{n-1} \\ m_n \end{pmatrix}=\begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \\ c_n \end{pmatrix} \tag{8}$$

将得出的 m_i 代入式(2), 即得三次样条插值函数 $\hat{s}(t)$。

可以用多项式插值、分段线性(或二次)插值等方法得到 $\hat{s}(t)$, 详见文献[3]。

2　推估误差

定理[3]　设 $s(t)$是$[a, b]$上二次连续可微的函数, 在$[a, b]$上以 $a=t_0<t_1<\cdots t_{n-1}<t_n=b$ 为节点的三次插值样条函数为$\hat{s}(t)$, 则对 $\forall t\in[t_{i-1}, t_i]$, 有:

$$|s(t)-\hat{s}(t)|\leqslant\frac{1}{2}|t_i-t_{i-1}|\max_t|s''(t)| \tag{9}$$

引理　设$s(t)$是$[a, b]$上二次连续可微的函数, 在$[a, b]$上以 $a=t_0<t_1<\cdots t_{n-1}<t_n=b$ 为节点, 过点$\{(t_i, s(t_i))\}_{i=0}^n$的三次插值样条函数为$\hat{s}(t)$, 过点$\{(t_i, \hat{s}(t_i))\}_{i=0}^n$的三次插值样条函数为$\hat{s}(t)$, 则对 $\forall t\in[t_{i-1}, t_i]$, 有:

$$\begin{aligned}|s(t)-\hat{s}(t)|\leqslant{}&|t_i-t_{i-1}|\cdot|(\hat{m}_i-m_i)-(\hat{m}_{i-1}-m_i)|+3|(\hat{s}(t_i)-s(t_i))+(\hat{s}(t_{i-1})\\&-s(t_{i-1}))|+\frac{1}{2}|t_i-t_{i-1}|\max_t|s''(t)|\end{aligned} \tag{10}$$

式中, m_i 是样条函数过点$\{(t_i, s(t_i))\}_{i=0}^n$; $\hat{m}_i$ 是样条函数过点$\{(t_i, \hat{s}(t_i))\}_{i=0}^n$。$m_i$ 与 $\hat{m}_i$ 由式(8)可得到。

证明　先对$\{(t_i, s(t_i))\}_{i=0}^n$及$\{(t_i, \hat{s}(t_i))\}_{i=0}^n$运用式(2)可得:

$$\begin{aligned}|\hat{s}(t)-\hat{s}(t)|={}&\left|(\hat{m}_{i-1}-m_{i-1})\frac{(t_i-t)^2(t-t_{i-1})}{h_i^2}-(\hat{m}_i-m_i)\frac{(t-t_{i-1})^2(t_i-t)}{h_i^2}\right|+\\&\left|(\hat{s}(t_{i-1})-s(t_{i-1}))\frac{(t_i-t)^2(2(t-t_{i-1})+h_i)}{h_i^3}+\right.\\&\left.(\hat{s}(t_i)-s(t_i))\frac{(t-t_{i-1})^2(2(t_i-t)+h_i)}{h_i^3}\right|,\\&t\in[t_{i-1}, t_i],\quad i=1, 2, \cdots, n\end{aligned}$$

于是, $|\hat{s}(t)-\hat{s}(t)|\leqslant|t_i-t_{i-1}|\cdot|(\hat{m}_i-m_i)-(\hat{m}_{i-1}-m_{i-1})|+3|(\hat{s}(t_i)-s(t_i))+(\hat{s}(t_{i-1})-s(t_{i-1}))|$ (11)

从而由式(11)及引理易得:

$$|s(t)-\hat{s}(t)|\leqslant|\hat{s}(t)-\hat{s}(t)|+|\hat{s}(t)-s(t)|\leqslant|t_i-t_{i-1}|\cdot|(\hat{m}_i-m_i)-(\hat{m}_{i-1}-m_{i-1})|+3|(\hat{s}(t_i)-s(t_i))+(\hat{s}(t_{i-1})-s(t_{i-1}))|+\frac{1}{2}|t_i-t_{i-1}|\max_t|s''(t)|$$

当 $s(t_i)$与其估计值 $\hat{s}(t_i)$的差别越小和节点越多(即观测值越多)时, $s(t)$的估值 $\hat{s}(t)$就越精确。这符合实际情况, 从而说明了引理的正确性和实际意义, 能够指导实际应用。

3　算　例

模拟半参数模型[1] $\boldsymbol{Y}=\boldsymbol{X}\beta+\boldsymbol{S}+\boldsymbol{e}$, 令 $\boldsymbol{X}=(x_{ij})_{100\times2}$, $\beta=(2, 3)^{\mathrm{T}}$, $x_{i1}=i/20$, $x_{i2}=(i/20)^2(i=1, 2, \cdots, 100)$, $\boldsymbol{S}=(s(t_1), \cdots, s(t_{100}))^{\mathrm{T}}$, $s(t_i)=10\sin(t_i)$, $t_i=\frac{2(i-1)\pi}{100}$。

取 $\boldsymbol{P}=\boldsymbol{I}$, $\alpha=0.05$, $\boldsymbol{R}=\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$, 其中矩阵[4]

$$\boldsymbol{G}=\begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \end{pmatrix}_{n-1,\,n}$$

当取观测值的真值 $\boldsymbol{Y}=\boldsymbol{X}\beta+\boldsymbol{S}$ 时, 由文献[1]的估计公式, 经计算可得 $\hat{s}$, 并可得 $M=(m_1, m_2, \cdots, m_{100})$, 即

$$M=(m_1, m_2, \cdots, m_{100})=$$

(9.238 3 9.560 1 9.769 1 9.623 8 9.509 1 9.334 6 9.130 2 8.885 7 8.607 28.294 7
7.946 8 7.569 1 7.159 9 6.720 1 6.257 7 5.768 0 5.252 9 4.718 8 4.166 43.596 7
3.009 1 2.414 9 1.807 8 1.192 3 0.575 5 −0.047 1 −0.669 0 −1.287 5 −1.901−42.508 5
−3.104 4 −3.690 3 −4.260 1 −4.813 6 −5.348 1 −5.860 4 −6.351 8 −6.814 9 −7.254−87.662 8
−8.040 3 −8.390 0 −8.700 3 −8.981 2 −9.223 5 −9.429 5 −9.599 8 −9.729 6 −9.823−09.873 7
−9.887 6 −9.862 3 −9.797 0 −9.691 9 −9.552 0 −9.357 6 −9.138 6 −8.896 0 −8.610−28.282 0
−7.913 3 −7.524 8 −7.103 2 −6.652 2 −6.175 5 −5.673 3 −5.146 1 −4.600 7 −4.034−13.451 2
−2.854 3 −2.243 9 −1.625 3 −0.999 9 −0.367 2 0.267 5 0.901 2 1.532 6 2.159 52.778 1
3.386 8 3.985 8 4.567 0 5.133 7 5.680 4 6.207 0 6.709 9 7.184 7 7.637 78.058 2
8.449 0 8.809 6 9.134 6 9.426 6 9.681 0 9.904 0 10.070 6 10.269 8 10.099 29.790 0)

于是由式(2),可得三次样条插值分段函数 $\hat{s}(t)$为:

$$\hat{s}(t)=m_{i-1}\frac{(t_i-t)^2(t-t_{i-1})}{h_i^2}-m_i\frac{(t-t_{i-1})^2(t_i-t)}{h_i^2}+s_{i-1}\frac{(t_i-t)^2(2(t-t_{i-1})+h_i)}{h_i^3}+$$

$$s_i\frac{(t-t_{i-1})^2(2(t_i-t)+h_i)}{h_i^3},\quad t\in[t_{i-1},t_i],h_i=\frac{\pi}{50},\quad i=2,\cdots,100$$

由此函数式就可以求出 $t\in[0,1.98\pi]$ 中的任何值。

图 1 中横轴为 t,纵轴为 $s(t)$的真值(用光滑曲线表示)、插值点(用“°”表示)及观测值(用“*”表示)。从图 1 可看出,插值曲线与非参数函数 $s(t)$曲线非常接近,从而说明了该方法的有效性。

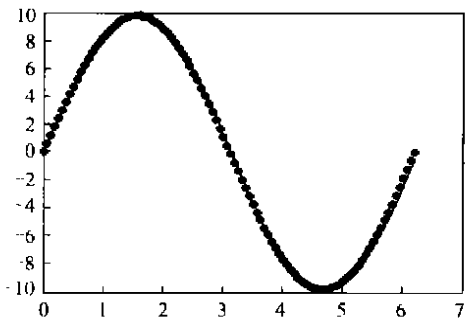

图 1 $s(t)$的真值、插值点及观测值

Fig. 1 Real Value, Interpolation Points and Observation Value of $s(t)$

当取有误差的观测值时,同样可计算得 M 及三次样条插值函数 $\hat{s}(t)$,参见图 2。图 2 中横轴为 t,纵轴为 $s(t)$的真值(用光滑曲线表示)、插值(用粗糙曲线表示)及观测值(用“*”表示)。

从图 1 和图 2 可看出,插值曲线与非参数函数 $s(t)$曲线的接近程度随观测值精度的提高而提高。

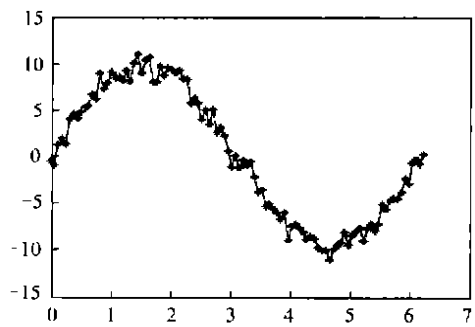

图 2 $s(t)$的真值、插值及观测值

Fig. 2 Real Value, Interpolation and Observation Value of $s(t)$

参 考 文 献

1 孙海燕,吴 云. 半参数回归与模型精化. 武汉大学学报·信息科学版, 2002, 27(2): 172~174

2 刘大杰,陶本藻. 实用测量数据处理方法. 北京:测绘出版社, 2000

3 陈明逵,凌永祥. 计算方法教程(工程类). 西安:西安交通大学出版社, 1992

4 Fischer B, Hegland M. Collocation, Filtering and Nonparametric Regression. ZfV, 1999(1): 17~24

5 郝红伟. Matlab 6实例教程. 北京:中国电力出版社, 2001

第一作者简介:胡宏昌,讲师,博士生。现从事测量数据处理理论及其应用的研究。代表成果:p-范分布参数σ的估计。

E-mail: retutome@yeah.net

Deducing and Estimating Nonparametric Signal in Semiparametric Regression Model

——Method of Cubic Splines Interpolation

HU Hongchang[1,2]　*SUN Haiyan*[1]

(1　School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2　Department of Mathematics, Hubei Normal University, Shenjiaying, Huangshi 435002, China)

Abstract: This paper considers the semiparametric regression model $y_i = \boldsymbol{X}_i^{\mathrm{T}}\beta + s(t_i) + e_i$ (for $i = 1, 2, \cdots, n$). Where $s_i = s(t_i)$ denotes the nonparametric signal of the observation and y_i a number relating to the observation at t_i, $\boldsymbol{X}_i \in R^p$ ($n > p$), $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ is parameter vector with p denoting the number of parameters, e_i denotes the noise and is assumed to be independently $N(0, \sigma_i^2)$ distributed.

Key words: semiparametric regression model; cubic spline function; deduce and estimate

About the first author: HU Hongchang, lecturer, Ph. D candidate. He is mainly engaged in the research on the theory and application of surveying data processing. His typical achievement is parameter σ estimation of p \| norm distribution.

E-mail: retutome@yeah.net

(责任编辑: 光远)

(上接第 637 页)

Stochastic Poisson Integral of Dirichlet Problem for Gravity Field

DENG Bo[1]　*ZHU Zhuowen*[1]　*LU Zhong*[2]

(1　Key Laboratory of Geospace Environment and Geodesy, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2　Department of Civil Engineering, The University of Hong Kong, Pokfulan Road, Hong Kong)

Abstract: In order to solve the boundary value problem with chaos or complicated boundary dates, such as singularity etc., it is necessary to establish the model of stochastic partial differential equation for the gravity field. This paper first constructs the conception of stochastic Sobolev spaces $H_2^2(\overline{\Omega})$, then gives the stochastic Poisson integral as a generalized stochastic functional, and compares the relationships between stochastic model and determined model of Poisson integral, and indicates that determined model is just one of the special situations of stochastic Poisson integral.

Key words: gravity field; dirichlet problem; stochastic Poisson integral

About the first author: DENG Bo, lecturer, Ph. D candidate. His major research orientation is the theory of stochastic model of geodetic boundary value problems.

(责任编辑: 平子)