DOI:CNKI:50-1165/N.20111110.1503.009

全局连续的分段最小二乘曲线拟合方法*

侯超钧,曾艳姗,吴东庆,杨志伟
(仲恺农业工程学院 计算科学学院,广州 510225)

摘要:针对传统最小二乘的曲线拟合方法不适合单一应用在具有复杂结构的试验观测数据中,本文提出一种满足全局连续性约束的多分段区间的最小二乘数据拟合方法。通过把每个相邻分段点上要求拟合连续的约束条件转化成一个矩阵等式 $Z\alpha=0$,建立一个只包含线性等式约束的最小二乘模型 $\min\limits_{\alpha}\|X\alpha-y\|^2$,最后通过应用拉格朗日的乘数方法推导出最小二乘解 α。本文的拟合方法在分段点上具有良好的拟合效果并满足全局连续,模型系数求解具有简单的显式表达式,易于编程数值计算。

关键词:曲线拟合;最小二乘;分段拟合;拉格朗日乘数法
中图分类号:O241.2 **文献标志码**:A **文章编号**:1672-6693(2011)06-0044-05

在工程实践与科学实验中,常常需要从一组带噪声的试验观测数据 (x_i,y_i),$i=1,2,\cdots,n$ 中找出自变量 x 与因变量 y 之间隐含的函数关系,一般采用数据拟合的办法来产生近似函数 $y=f(x)$。其中,线性最小二乘方法是解决曲线拟合问题的常用方法,通过采用一组简单合适的、线性无关的基函数来逼近试验数据,可以有效得出总体经验误差最小的拟合函数 $f(x)$[1]。

对于具有多个显著局部特征的试验观测数据,在全局定义域上应用一组基函数,将会得到性能较差的拟合函数。一种解决手段是采用分段曲线拟合,在每段区间上进行局部最小二乘拟合。然而,拟合函数在区间分段点上不一定连续,在相邻区间边界附近的拟合效果不理想[2-4]。其中,文献[5-6]提出多项式基函数的全局连续拟合方法,但只限于 2 个分段区间。文献[7]给出多分段区间、全局连续的曲线拟合方法,但基函数只限于一次多项式。文献[8]提出分段区间重合的拟合方法,由每 4 个数据点决定一个三次曲线,但分区间太密,不适用于密集的数据拟合。目前的研究方法还没有一个可以描述全局连续的分段曲线拟合的有效模型,以及相应的有效求解方法。

本文提出一种全局连续的多区间最小二乘曲线拟合方法,把全局连续性约束的分段拟合问题转化为带等式约束的误差最小化模型,通过拉格朗日乘数法推导出模型的最小二乘回归系数。本文方法在实际应用中具有以下优点:1)分段点上具有良好拟合效果,符合实际数据的连续性变化要求,从而避免拟合函数在分段点上的二义性;2)模型的回归系数求解具有简单的显式表达式,易于数值计算;3)方法对分段区间个数与观测数据规模的限制较小,在各个分段区间上可以采用不同的基函数组,容易编程实现。

1 两段区间的全局连续的最小二乘模型

假设有试验观测数据 (x_i,y_i),$i=1,2,\cdots,n$,可分为两个数据集 $S_1=\{(x_i^{(1)},y_i^{(1)})_{i=1}^{n_1}\}$ 与 $S_2=\{(x_i^{(2)},y_i^{(2)})_{i=1}^{n_2}\}$,满足 $x_i^{(1)}\leqslant x_0$,$1\leqslant i\leqslant n_1$ 与 $x_i^{(2)}\geqslant x_0$,$1\leqslant i\leqslant n_2$,$n_1+n_2=n$,其中 x_0 是两数据集的分割点。需要确定的拟合函数 $f(x)$ 的形式为

$$f(x)=\begin{cases}f_1(x)=\sum\limits_{j=1}^{m_1}\alpha_j^{(1)}h_j^{(1)}(x),x\leqslant x_0\\ f_2(x)=\sum\limits_{j=1}^{m_2}\alpha_j^{(2)}h_j^{(2)}(x),x\geqslant x_0\end{cases}\tag{1}$$

* 收稿日期:2011-05-22 修回日期:2011-07-12 网络出版时间:2011-11-10 15:03
资助项目:仲恺农业工程学院科研基金(No. G3100038)
作者简介:侯超钧,男,讲师,博士,研究方向为智能信息处理。
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20111110.1503.201106.44_009.html

其中 $f_1(x_0)=f_2(x_0)$,$\{h_j^{(k)}\}_{j=1}^{m_k}$ 是给定在 S_k 上一组线性无关的基函数,m_k 是 S_k 上基函数的个数,基函数一般可取简单形式的函数,如 $1,x,x^2,\sin x$ 等,不同数据集上可以选取不相同的基函数组。$\{\alpha_j^{(k)}\}_{j=1}^{m_k}$ 则是待确定的回归系数,则使总体拟合误差最小且在 x_0 上连续的两分段最小二乘回归模型为

$$\min_{\alpha_j^{(1)},\alpha_j^{(2)}} \sum_{i=1}^{n_1}\left(f_1(x_i^{(1)})-y_i^{(1)}\right)^2+\sum_{i=1}^{n_2}\left(f_2(x_i^{(2)})-y_i^{(2)}\right)^2$$
$$\text{s.t.}\quad f_1(x_0)=f_2(x_0) \tag{2}$$

令行向量函数 $h^{(k)}(x)=[h_1^{(k)}(x),h_2^{(k)}(x),\cdots,h_{m_k}^{(k)}(x)]$,$k=1,2$

$$X_k=\begin{bmatrix} h^{(k)}(x_1^{(k)}) \\ h^{(k)}(x_2^{(k)}) \\ \square \\ h^{(k)}(x_{n_k}^{(k)}) \end{bmatrix},y_k=\begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \square \\ y_{n_k}^{(k)} \end{bmatrix},\alpha_k=\begin{bmatrix} \alpha_1^{(k)} \\ \alpha_2^{(k)} \\ \square \\ \alpha_{m_k}^{(k)} \end{bmatrix}.$$

则模型(2)可用矩阵向量形式表达:

$$\min_{\alpha_1,\alpha_2} \|X_1\alpha_1-y_1\|^2+\|X_2\alpha_2-y_2\|^2$$
$$\text{s.t.}\quad z_1^{\mathrm{T}}\alpha_1=z_2^{\mathrm{T}}\alpha_2 \tag{3}$$

其中 $z_k\triangleq[h^{(k)}(x_0)]^{\mathrm{T}}$。拉格朗日函数为 $L(\alpha_1,\alpha_2,\lambda)=\|X_1\alpha_1-y_1\|^2+\|X_2\alpha_2-y_2\|^2+2\lambda(z_1^{\mathrm{T}}\alpha_1-z_2^{\mathrm{T}}\alpha_2)$,由求多元函数极值的必要条件[9],有

$$\begin{cases} \dfrac{\partial L}{\partial\alpha_1}=2X_1^{\mathrm{T}}(X_1\alpha_1-y_1)+2\lambda z_1=0 \\ \dfrac{\partial L}{\partial\alpha_2}=2X_2^{\mathrm{T}}(X_2\alpha_2-y_2)-2\lambda z_2=0 \\ \dfrac{\partial L}{\partial\lambda}=z_1^{\mathrm{T}}\alpha_1-z_2^{\mathrm{T}}\alpha_2=0 \end{cases} \tag{4}$$

由(4)式的头两个方程可整理得

$$\alpha_1=\hat{\alpha}_1-\lambda(X_1^{\mathrm{T}}X_1)^{-1}z_1,\alpha_2=\hat{\alpha}_2+\lambda(X_2^{\mathrm{T}}X_2)^{-1}z_2 \tag{5}$$

其中 $\hat{\alpha}_k\triangleq(X_k^{\mathrm{T}}X_k)^{-1}X_k^{\mathrm{T}}y_k$,$k=1,2$。把(5)式中 α_1 与 α_2 的表达式代入等式 $z_1^{\mathrm{T}}\alpha_1-z_2^{\mathrm{T}}\alpha_2=0$,可解得

$$\lambda=\frac{z_1^{\mathrm{T}}\hat{\alpha}_1-z_2^{\mathrm{T}}\hat{\alpha}_2}{z_1^{\mathrm{T}}(X_1^{\mathrm{T}}X_1)^{-1}z_1+z_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)^{-1}z_2} \tag{6}$$

然后把 λ 回代入(5)式,得到最小二乘回归系数 α_1 与 α_2。一般地,实际数据集 S_1 与 S_2 都可以保证 $(X_1^{\mathrm{T}}X_1)^{-1}$与$(X_2^{\mathrm{T}}X_2)^{-1}$是非退化的,从而 α_1 与 α_2 有唯一解,即拟合函数 $f(x)$ 是有效的。以下将以矩阵向量的形式推广到一般的多分段全局连续的最小二乘模型。

2 多分段区间的全局连续的最小二乘模型

考虑 K 个数据集 $S_k=\{(x_i^{(k)},y_i^{(k)})_{i=1}^{n_k}\}$,$k=1,2,\cdots,K$,它们在实数轴上分为 K 个区间,S_k 中的数据满足 $x_{k-1}\leqslant x_i^{(k)}\leqslant x_k$,$1\leqslant i\leqslant n_k$,其中 x_k 是 S_k 与 S_{k+1} 的分段点,约定 $x_0=-\infty$,$x_K=+\infty$。令 $y=[y_1^{\mathrm{T}},y_2^{\mathrm{T}},\cdots,y_K^{\mathrm{T}}]^{\mathrm{T}}$,$\alpha=[\alpha_1^{\mathrm{T}},\alpha_2^{\mathrm{T}},\cdots,\alpha_K^{\mathrm{T}}]^{\mathrm{T}}$

$$X=\begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \square & \square & & \square \\ 0 & 0 & \cdots & X_K \end{bmatrix},Z=\begin{bmatrix} h^{(1)}(x_1) & -h^{(2)}(x_1) & 0 & \cdots & 0 & 0 \\ 0 & h^{(2)}(x_2) & -h^{(3)}(x_2) & \cdots & 0 & 0 \\ \square & \square & \square & & \square & \square \\ 0 & 0 & 0 & \cdots & -h^{(K-1)}(x_{k-2}) & 0 \\ 0 & 0 & 0 & \cdots & h^{(K-1)}(x_{k-1}) & -h^{(K)}(x_{k-1}) \end{bmatrix} \tag{7}$$

其中 X 是 $\sum_{k=1}^{K}n_k\times\sum_{k=1}^{K}m_k$ 的对角块矩阵,X_k 是 S_k 中的数据根据基函数组 $\{h_j^{(k)}\}_{j=1}^{m_k}$ 所生成的矩阵,Z 是 $(K-1)\times\sum_{k=1}^{K}m_k$ 的块矩阵。则多分段区间的全局连续的最小二乘回归模型可以表达成

$$\min_{\alpha} \| X\alpha - y \|^2$$
$$\text{s.t. } Z\alpha = 0 \tag{8}$$

其中等式约束描述了拟合函数$f(x)$在每个分段点$\{v_k\}$处的连续性,即$f_k(x_k) = f_{k+1}(x_k)$,$k = 1,2,\cdots,K-1$。对(8)式建立拉格朗日函数

$$L(\alpha,\lambda) = \| X\alpha - y \|^2 + 2\lambda^{\mathrm{T}} Z\alpha \tag{9}$$

其中λ是长度为$K-1$的列向量。对$L(\alpha,\lambda)$求偏导得

$$\begin{cases} \dfrac{\partial L}{\partial \alpha} = 2X^{\mathrm{T}}(X\alpha - y) + 2Z^{\mathrm{T}}\lambda = 0 \\ \dfrac{\partial L}{\partial \lambda} = Z\alpha = 0 \end{cases} \tag{10}$$

由(10)式的第1个等式可得

$$\alpha = \hat{\alpha} - (X^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}\lambda \tag{11}$$

其中$\hat{\alpha} \triangleq (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$,把$\alpha$的表达式代入(10)式第2个等式,可解得

$$\lambda = [Z(X^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}]^{-1}Z\hat{\alpha} \tag{12}$$

由此把λ表达式代入(11)式即得最小二乘回归系数α。

如果在模型(8)中没有等式约束,即不要求$f(x)$在分割点$\{x_k\}$上连续,此时$\lambda = 0$,公式(11)的α就退化为$\hat{\alpha}$,由X的块对角性,得$\hat{\alpha}_k = (X_k^{\mathrm{T}}X_k)^{-1}X_k^{\mathrm{T}}y_k$,与独立求解每个分段上曲线拟合函数的情形一致,$\alpha$可看成由分段上独立最小二乘系数$\hat{\alpha}$再加上连续性约束的修正。相对于其他分段区间回归的研究方法,(11)式具有简单的显式表达式,易于数值计算。

在(11)、(12)式的矩阵$X^{\mathrm{T}}X$的维数一般并不高,只与各分段区间使用基函数组的总个数有关,在密集数据的最小二乘应用中,通常都远远小于观测数据点个数,在实际数值计算中,还可以根据$X^{\mathrm{T}}X$与其逆矩阵都是对称块对角矩阵的性质,可以进一步简化数值计算复杂度,这将是进一步的研究内容。

3 数值实验

实验采用函数$y = 12\mathrm{e}^{-2x}$在$[0.5,16]$区间上均匀间隔生成32个样本点,并且加入了高斯噪声$N(0,0.25)$组成观测数据(x_i,y_i),$i = 1,2,\cdots,32$。由图1看出有两段曲线趋势不同的区间,这里取分段点$x_0 = 4$,在两个区间上分别采用线性函数进行拟合,即$h^{(1)}(x) = h^{(2)}(x) = [1,x]$。图2显示了对每个分段区间进行独立的最小二乘方法,由于没有在分段点上施加连续性约束,两个分段函数在分段点处出现较大差异。图3则显示了本文采用连续性约束的分段最小二乘方法,利用(5)式计算得$\alpha_1 = [-0.441,2.257]^{\mathrm{T}}$,$\alpha_2 = [7.765,0.206]^{\mathrm{T}}$,即分段拟合函数为

$$f(x) = \begin{cases} f_1(x) = h^{(1)}(x) \cdot \alpha_1 = -0.441 + 2.257x, & x \leqslant x_0 \\ f_2(x) = h^{(2)}(x) \cdot \alpha_2 = 7.765 + 0.206x, & x \geqslant x_0 \end{cases} \tag{13}$$

显然相比图2,采用(13)的拟合函数的更符合实际数据的连续性变化趋势。

接着考虑以下定义的分段函数

$$g(x) = \begin{cases} 0.6(x+1), & -3 \leqslant x \leqslant -1 \\ -x^2 + 1, & -1 \leqslant x \leqslant 1 \\ -0.6(x-1), & 1 \leqslant x \leqslant 3 \end{cases} \tag{14}$$

分别在$[-3,-1]$,$[-1,1]$与$[1,3]$内各生成间隔为0.2的10个样本点,并且加入了高斯噪声$N(0,0.125^2)$组成3个数据集S_1,S_2,S_3,见图4。该实验采用3个分段区间,两个分段点$x_1 = -1$,$x_2 = 1$,其中S_1与S_3采用线性函数拟合,S_2采用二次函数拟合,即$h^{(1)}(x) = h^{(3)}(x) = [1,x]$,$h^{(2)}(x) = [1,x,x^2]$。由图5看到采用独立区间的曲线拟合方法在分段点处有明显的不连续,有较大的数值差异,这对于在分段点附近的拟合值存在较大的二义性。图6是使用本文具有分段连续约束的曲线拟合方法的示意图,由(11)式可以计算得$\alpha_1 = [0.604,0.621]^{\mathrm{T}}$,$\alpha_2 = [0.977,0.090,-0.904]^{\mathrm{T}}$,$\alpha_3 = [0.954,-0.791]^{\mathrm{T}}$,即分段拟合函数表

达为

$$f(x)=\begin{cases} f_1(x)=h^{(1)}(x)\cdot\alpha_1=0.604+0.621x, & -3\leqslant x\leqslant -1 \\ f_2(x)=h^{(2)}(x)\cdot\alpha_2=0.977+0.09x-0.904x^2, & -1\leqslant x\leqslant 1 \\ f_3(x)=h^{(3)}(x)\cdot\alpha_3=0.954-0.791x, & 1\leqslant x\leqslant 3 \end{cases} \tag{15}$$

由图 6 看出,采用(15)式的分段拟合函数在分段点上是连续的,更符合实际数据的连续性变化趋势。此外,采用对分段区间进行独立最小二乘方法的总体拟合的均方误差为 $MSE=0.115\ 1$,而(15)式计算得 $MSE=0.125\ 1$,与样本点的噪声水平相当。可见,增加了连续性约束后 MSE 的变化不大,是可接受的。

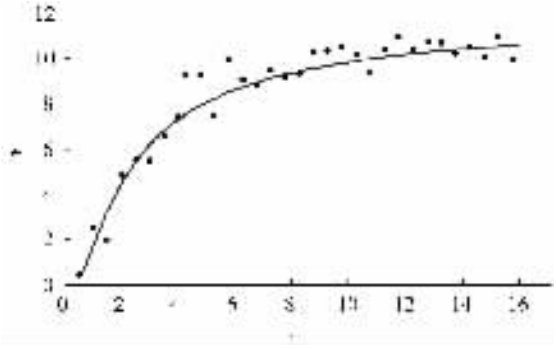

图 1　两分段的数据样本点

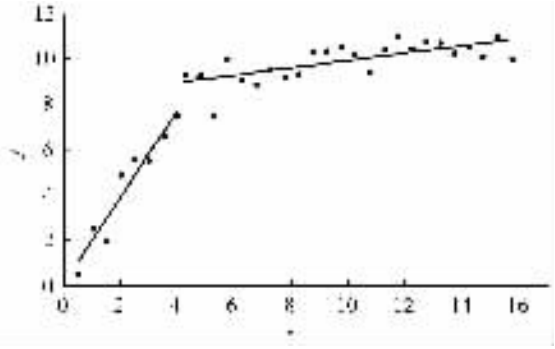

图 2　两分段的独立曲线拟合

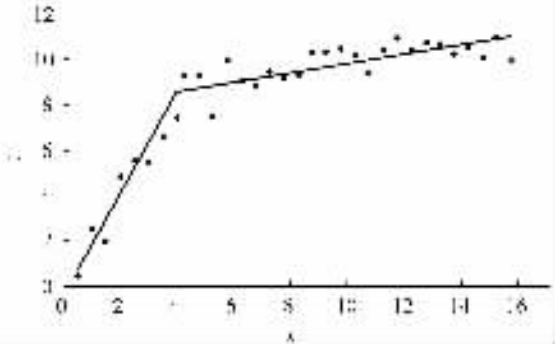

图 3　带连续约束的分段曲线拟合

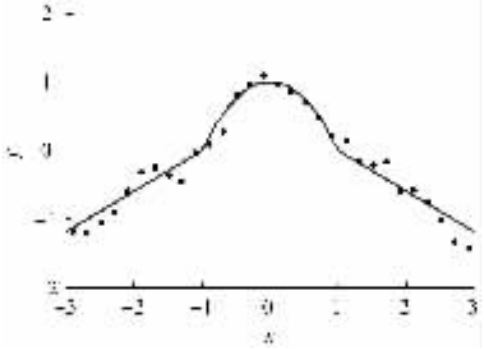

图 4　三分段的数据样本点

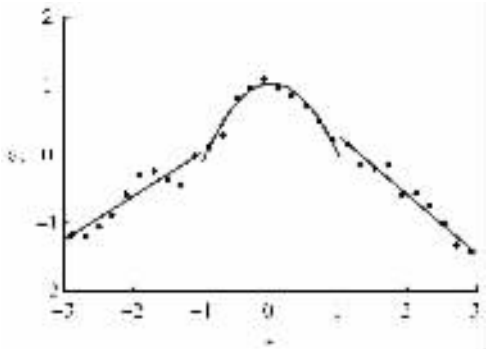

图 5　三分段的独立曲线拟合

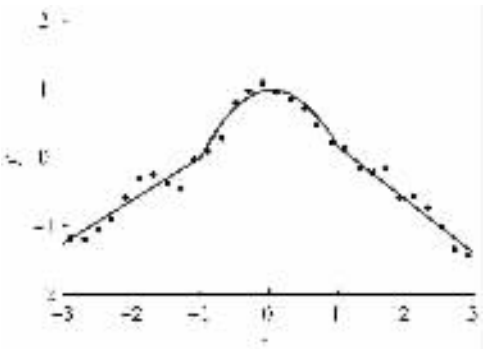

图 6　三分段带连续约束的曲线拟合

4　结论

本文提出一种带全局连续约束的多区间的最小二乘曲线拟合方法,在每个分段曲线上可以采用不同的基函数组。笔者把多分段曲线拟合的全局连续性问题构造成一个只带等式约束的最小二乘模型,并应用拉格朗日的乘数方法求解矩阵向量形式的最优化问题。本方法具有简单的计算表达式,能满足拟合函数在分段点上的连续性,数值实验表明了该方法具有较好的拟合效果。

参考文献：

[1] 李庆扬,王能超,易大义. 数值分析[M]. 武汉 华中科技出版社,2001.

[2] 陈廷楠,张继顺. 样条函数在估算飞机气动参数中的应用[J]. 飞行力学,1995,14(3):61-65.

[3] 万辉,魏延. 一种改进的最小二乘支持向量机算法[J]. 重庆师范大学学报:自然科学版,2010,27(4):69-73.

[4] 黄敬频. 矩阵方程组[A_1XB_1,A_2XB_2] = [C,D]的最小二乘解[J]. 四川师范大学学报:自然科学版,2003(4):370-372.

[5] 刘晓莉,陈春梅. 基于最小二乘原理的分段曲线拟合法[J]. 伊犁教育学院学报,2004,17(3):132-134.

[6] 尹花兵,帅立国,姜昌金,等. 基于分段最小二乘方法的钢材无损检测分选仪[J]. 机械工程与自动化,2008(1):105-107.

[7] Gluss B. An alternative method for continuous line segment curve-fitting[J]. Information and Control,1964,7(2):200-206.

[8] 蔡山,张浩,陈洪辉,等. 基于最小二乘法的分段三次曲线拟合方法研究[J]. 科学技术与工程,2007,7(3):352-355.

[9] Boyd S, Vandenberghe L. Convex optimization[M]. UK: Cambridge University Press,2004.

Global Continuous Curve Fitting Method of Piecewise Least Square Fitting with Global Continuity

HOU Chao-jun, CENG Yan-shan, WU Dong-qing, YANG Zhi-wei

(School of Computational Science, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China)

Abstract: The application of traditional least square method is limited when the experiment data are of heterogeneous structure. This paper presents a novel constrained least square method for solving the piecewise local curve fitting problem with global continuity constraint. In particular, the continuity constraint among the segment points was converted to the matrix equality $Z\alpha = 0$. Therefore, the least square model $\min_{\alpha} \| X\alpha - y \|^2$ was proposed with the linear equality constraint to address the problem. Here, the solution of the least square model was seriously derived in a simple explicit form via the Lagrange multiplier method, which can be easily programmed in numerical calculation. The experimental results show that the method proposed here provided a" best "fit for the data and the global continuity on the segment points.

Key words: curve fitting; least square; piecewise fitting; Lagrange multiplier method

(责任编辑 游中胜)