

## 极差调节的局部惩罚样条回归方法\*

李 猛 杨联强\* 江 坤 何声娴

(安徽大学数学科学学院, 合肥, 230601)

**摘 要:** 在惩罚样条回归模型中, 根据截断幂基函数系数的直观意义, 以结点两边数据点极差的线性递减函数作为局部惩罚权重, 构造了一种新的局部惩罚样条回归模型. 不同于整体惩罚样条, 该方法使得当数据点集在局部具有较大的波动性时, 能给予拟合曲线较小的惩罚, 从而能更好地控制曲线在拟合优度与光滑度之间的平衡. 模拟结果显示, 当数据具有空间异质性时, 采用该方法的回归模型相比整体惩罚模型有更好的信息准则得分.

**关键词:** 整体惩罚样条; 局部惩罚样条; 异方差性; 极差

**中图分类号:** O212.7

## §1. 引 言

回归分析的任务是从已知数据点集 $\{(x_i, y_i), i = 1, 2, \dots, n\}$ 中得到响应变量 $y$ 与解释变量 $x$ 之间的函数关系<sup>[1]</sup>. 模型记作

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n. \quad (1)$$

通常假定 $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)' = \varepsilon \sim N(0, \sigma^2 I)$ , 求 $y = f(x) = E(y|x)$ .

回归模型总体上可分为两大类, 参数模型和非参数模型. 其中参数模型是假定函数 $f(x)$ 的形式已知, 而其中的参数未知, 所以模型的任务就转化为对参数的估计. 广泛应用的多元线性回归模型即假定 $f(x) = \mathbf{x}\boldsymbol{\beta}$ , 其中 $\mathbf{x} = (x_1, x_2, \dots, x_n)$ 为解释变量,  $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_n)'$ 为未知参数. 参数模型形式简洁、计算高效, 已建立完善的理论, 实际应用及其广泛<sup>[2]</sup>, 但其缺点是模型的拟合能力柔性不足, 而且可能会有模型设定错误的风险. 非参数方法不假定函数 $f(x)$ 的具体形式<sup>[3]</sup>, 而是用一些函数拟合工具来拟合 $f(x)$ , 常见的有核(kernel)回归、样条(spline)回归、局部多项式回归等<sup>[3,4]</sup>. 非参数方法的优点是拟合能力强, 但缺点是模型的计算和检验复杂, 且存在“维数诅咒”问题. 该方法是当前回归分析中非常活跃的研究领域之一<sup>[5,6]</sup>. 样条函数源于数值分析领域的研究工作, 被引入到统计学后在非参数回归中得到迅速发展和应用<sup>[3,4]</sup>, 光滑样条系列模型是典型代表<sup>[7]</sup>. 惩罚样条回归(penalized spline regression)首先由Eilers和Marx<sup>[8]</sup>提出, 其以 $B$ 样条为函数拟合工具,

\*国家自然科学基金天元基金(11026076)和安徽大学科研训练计划项目(J18520205)资助.

\*通讯作者, E-mail: yanglq@ahu.edu.cn.

本文2014年6月9日收到, 2014年11月6日收到修改稿.

在目标函数中加入样条系数二阶差分的和作为惩罚, 以此来控制曲线对数据点的过度拟合现象. Ruppert等<sup>[9]</sup>详细介绍了基于截断幂基的惩罚回归样条, 其惩罚项设置为截断幂基函数系数的平方和. 这两种代表性的惩罚样条方法近年来得到深入研究和广泛应用, 但在这两种方法中, 对样条系数的惩罚都是均匀惩罚, 或者称作整体惩罚, 即对所有位置的样条系数的惩罚权重是相同的, 没有考虑数据的空间异质性对惩罚的局部性要求. Ruppert和Carroll<sup>[10]</sup>给出了一种局部惩罚样条, 但该方法的局部惩罚参数的计算过于复杂, 理论上需要多维网格搜索来寻找最优惩罚参数值, 因此作者将其转化为多重一维网格搜索来寻找局部最优值, 且惩罚参数值没有直观解释.

本文给出一种直观的局部惩罚样条回归方法, 通过引入分段范围内数据的极差来调节惩罚权重, 使得该方法更适用于具有空间异质性的数据. 模型的求解计算简洁, 只需运用经典的岭回归技术, 而最优惩罚参数的选取也只需简单的一维网格搜索. 数据模拟结果显示该方法比整体惩罚样条回归具有明显的信息准则得分优势.

## §2. 整体惩罚样条回归简介

本节对基于截断幂基函数的整体惩罚样条作一简介<sup>[9]</sup>.

记 $p$ 次样条截断幂基函数为

$$\mathbf{x} = (1, x, x^2, \dots, x^p, (x - \kappa_1)_+^p, (x - \kappa_2)_+^p, \dots, (x - \kappa_K)_+^p),$$

其中 $\kappa_1 < \kappa_2 < \dots < \kappa_K$ 为选定的结点, 截断幂函数

$$(x - \kappa_i)_+^p = \begin{cases} 0, & x \leq \kappa_i; \\ (x - \kappa_i)^p, & x > \kappa_i. \end{cases}$$

设 $f(x) = \mathbf{x}\boldsymbol{\beta}$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_K)'$ 为样条系数. 对给定的数据集 $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , 记 $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ , 设计矩阵 $\mathbf{X} = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n)'$ , 其中 $\mathbf{x}_i = (1, x_i, x_i^2, \dots, x_i^p, (x_i - \kappa_1)_+^p, (x_i - \kappa_2)_+^p, \dots, (x_i - \kappa_K)_+^p)$ . 则模型(1)中参数的无惩罚普通最小二乘(OLS)解为

$$\hat{\boldsymbol{\beta}} = \arg \min \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (2)$$

显然, 由于目标函数只考虑了对数据的拟合优度, 而没有考虑拟合曲线的光滑优度, 因而拟合曲线 $\hat{f}(x) = \mathbf{x}\hat{\boldsymbol{\beta}}$ 在结点较多时存在对数据的过度拟合现象, 即曲线有过多不必要的局部震荡. 因此, 为了克服这种现象, 通常是在目标函数中加入粗糙度惩罚项, 进而参数的估计值为

$$\hat{\boldsymbol{\beta}} = \arg \min \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}\} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}'\mathbf{y}, \quad (3)$$

其中,  $\lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}$ 为对截断幂系数的惩罚,  $\lambda > 0$ 为惩罚参数(也称作光滑参数), 惩罚矩阵 $\mathbf{D} = \text{diag}(\mathbf{0}_{p+1}, \mathbf{1}_K)$ 是主对角线前 $p+1$ 个元素为0后 $K$ 个元素为1的对角矩阵. 因此, 该惩罚项的

设置中, 每个参数 $\beta_{p+1}, \beta_{p+2}, \dots, \beta_{p+K}$ 的惩罚权重是相等的, 总惩罚量由参数 $\lambda$ 来调节. 我们称这种方法为整体惩罚样条回归, 或者叫做均匀惩罚样条回归.

### §3. 局部惩罚样条回归

整体惩罚样条通过某种信息准则寻找最优惩罚参数 $\lambda$ , 使得拟合曲线在光滑度和拟合优度之间得到良好的平衡, 从而取得了很好的模型拟合效果. 但观察惩罚矩阵可知, 当数据集具有显著的空间异质性(例如异方差性)时, 对不同结点位置的系数施行不同权重的惩罚将是更合理的选择. 亦即考虑如下类型的惩罚矩阵

$$D(\alpha) = \text{diag}(0, 0, \dots, 0, \alpha(\kappa_1), \alpha(\kappa_2), \dots, \alpha(\kappa_K))$$

并如何合理选择惩罚权重 $\alpha(\kappa_1), \alpha(\kappa_2), \dots, \alpha(\kappa_K)$ 成为值得研究的问题. 理论上, 我们可以通过 $R^K$ 中全局搜索来寻找最优的 $\alpha(\kappa_1), \alpha(\kappa_2), \dots, \alpha(\kappa_K)$ 的取值, 但在结点个数较多时计算量太大, 因而几乎不可能实现. 为此, Ruppert等<sup>[9]</sup>通过逐步的一维网格搜索少数几个结点位置的 $\alpha(\kappa_i)$ , 其他位置的 $\alpha(\kappa_j)$ 通过对 $\alpha(\kappa_i)$ 的线性插值来实现, 显然该方法计算还是比较复杂, 而且也只是一种折中的处理方法.

观察样条函数 $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p + \beta_{p+1}(x - \kappa_1)_+^p + \beta_{p+2}(x - \kappa_2)_+^p + \dots + \beta_{p+K}(x - \kappa_K)_+^p$ 可知,  $f(x)$ 正是通过对多项式函数 $\beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p$ 在每个结点 $\kappa_i$ 处加入一个“扰动项” $\beta_{p+i}(x - \kappa_i)_+^p$ , 以此实现曲线的局部变化, 从而能柔性地去拟合数据的(混合模型框架下惩罚样条回归的求解即是基于此种思想, 该框架下将 $\beta_{p+i}(x - \kappa_i)_+^p$ 视作随机效应, 而将 $\beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p$ 视作固定效应<sup>[9]</sup>). 因此, 从直观意义上, 惩罚项 $\lambda \beta' D \beta$ 的设置, 正是通过控制各 $\beta_{p+i}$ 平方和的大小来控制这些“扰动项”作用的大小. 基于此直观意义, 当数据本身在某个结点附近具有较大的波动性时, 也应该允许拟合曲线在此位置有更多波动的自由, 因此在惩罚项中应给予较小的权重. 反之, 则应减少曲线在该处波动的自由, 即应加大惩罚项权重. 在此意义上, 我们给出如下直观的局部惩罚权重的设置, 令

$$\rho_k = \max\{|x_i - x_j| \mid x_i, x_j \in [\kappa_{k-1}, \kappa_{k+1}]\}, \quad k = 1, 2, \dots, K,$$

$$L = \max(\rho_k) + 1,$$

构造局部惩罚权重为 $\alpha(\kappa_k) = L - \rho_k, k = 1, 2, \dots, K$ . 即将每个结点处的局部惩罚权重设置为该结点左右相邻两区间内数据纵向极差的线性正值单调减函数. 此时参数拟合结果即为

$$\hat{\beta} = \arg \min \{\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \beta' D(\alpha) \beta\} = (\mathbf{X}' \mathbf{X} + \lambda D(\alpha))^{-1} \mathbf{X}' \mathbf{y}, \quad (4)$$

则 $\mathbf{y}$ 的拟合值 $\hat{\mathbf{y}} = \mathbf{X} \hat{\beta}$ .

另外, 本文结点数目和位置的选取应用了Ruppert和Carroll<sup>[10]</sup>推荐的原则. 使用样本值 $x_i, i = 1, 2, \dots, n$ 的等分位点作为结点, 即结点 $\kappa_k$ 为 $x_i$ 的第 $\kappa_k = (k + 1)/(K + 2)$ 个分位点.

惩罚参数 $\lambda$ 的选取可以根据多种信息准则来决定, 本文采用大多数文献所推荐的广义交叉验证(generalized-cross-validation, GCV)准则进行<sup>[4, 9, 10]</sup>, 其中

$$\text{GCV}(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\{1 - n^{-1} \text{Tr}(\mathbf{H}_\lambda)\}^2},$$

而 $\text{Tr}(\mathbf{H}_\lambda)$ 是 $\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D}(\alpha))^{-1}\mathbf{X}'$ 的迹,  $\lambda$ 是使得 $\text{GCV}(\lambda)$ 达到最小的取值, 可以通过一维网格搜索获得.

## §4. 模 拟

本节给出两个模拟实例, 以显示本文所提出的局部惩罚样条的拟合效果, 所有计算、作图、分析工作均在R3.02中完成.

**例 1**  $f(x) = 10(\sin x/x)$ ,  $y_i = f(x_i) + \varepsilon_i$ ,  $x_i \sim U[1, 20]$ ,  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $n = 600$ .  $\sigma^2$ 取了六个不同值, 从左至右逐渐增大. 一次随机试验结果的散点图、整体惩罚拟合曲线、局部惩罚拟合曲线见图1.

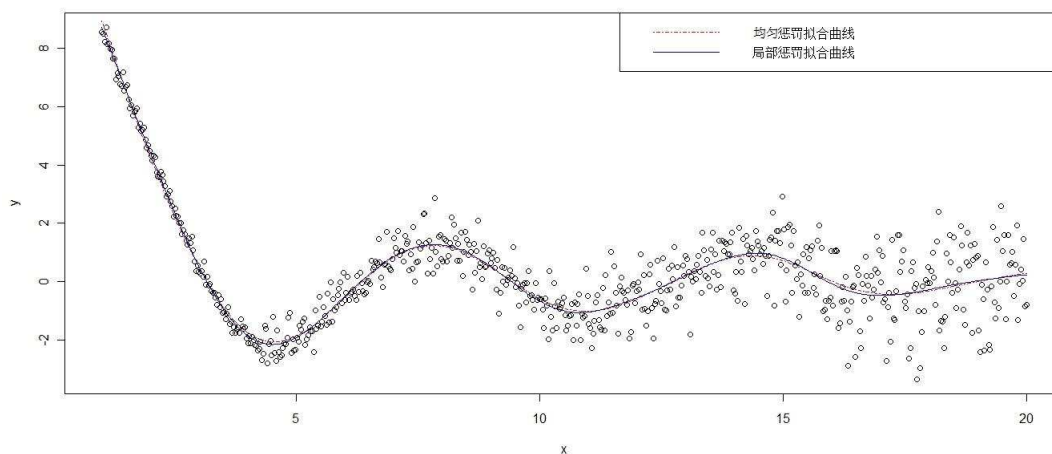


图1 散点、整体惩罚拟合曲线、极差调节的局部惩罚拟合曲线图

图2a采用整体惩罚样条回归, 最优惩罚参数 $\lambda = 10.150505$ , 各项信息准则得分为

$$C_{p1} = 438.7787, \quad \text{GCV}_1 = 438.9091, \quad \text{AIC}_1 = 6.083864, \quad \text{RMSE}_1 = 20.51942;$$

图2b采用局部惩罚样条回归, 最优惩罚参数 $\lambda = 6.043434$ , 各项信息准则得分为

$$C_{p2} = 438.4203, \quad \text{GCV}_2 = 438.5498, \quad \text{AIC}_2 = 6.083046, \quad \text{RMSE}_2 = 20.51178.$$

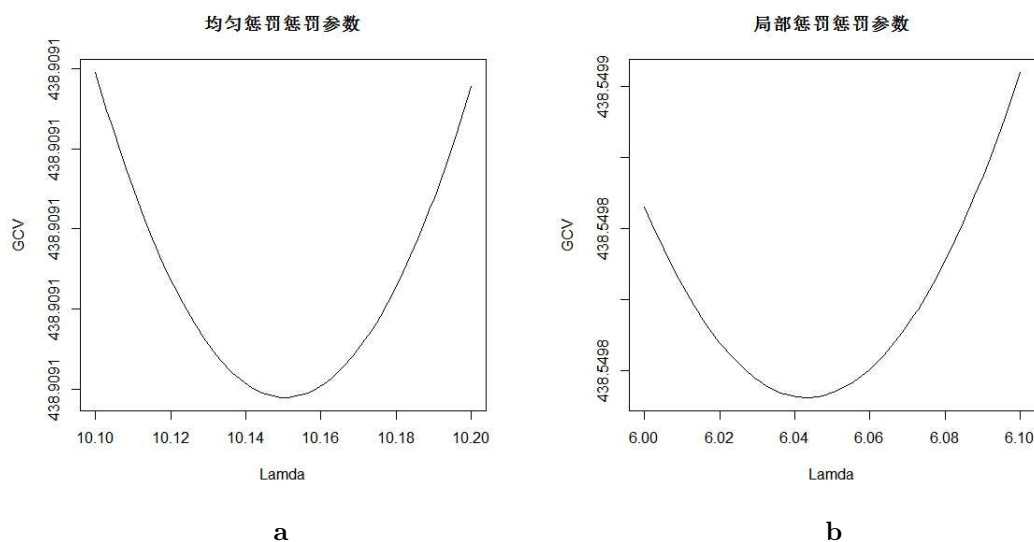


图2 惩罚参数 $\lambda$ 与 $GCV(\lambda)$ 函数关系图(a. 整体惩罚, b. 极差调节的局部惩罚)

同样可见, 在各项信息准则下, 极差调节的局部惩罚样条回归模型具有更好的表现.

**例 2** 令  $f(x) = 10\sqrt{x(1-x)}\sin((18\pi)/(0.8x+1))$ ,  $y_i = f(x_i) + \varepsilon_i$ ,  $x_i \sim U[0, 1]$ ,  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $n = 660$ .  $\sigma^2$ 取了六个不同值, 从左至右先增后减. 一次随机试验结果的散点图、整体惩罚拟合曲线、局部惩罚拟合曲线见图3.

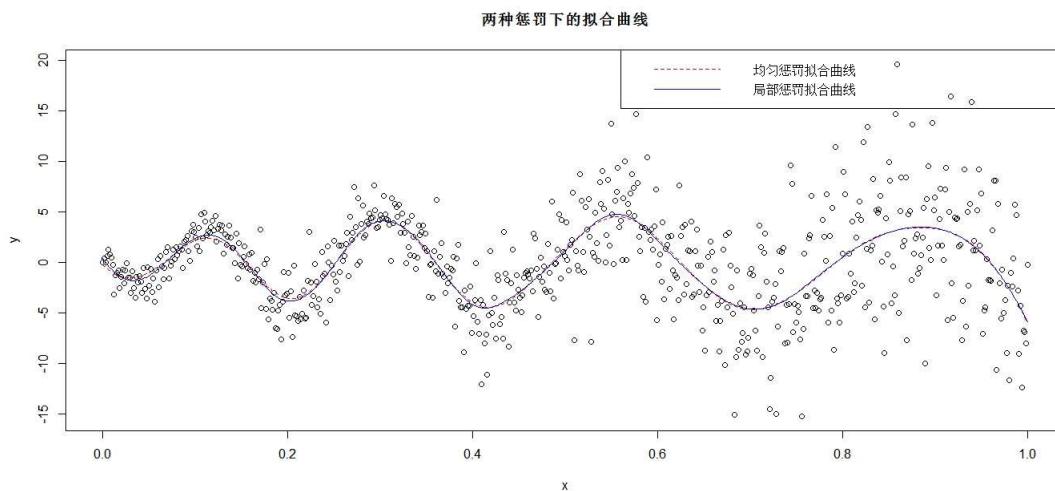


图3 散点、整体惩罚拟合曲线、极差调节的局部惩罚拟合曲线图

图4a采用整体惩罚样条回归, 最优惩罚参数  $\lambda = 1.565657 \times 10^{-5}$ , 各项信息准则得分为

$$C_{p1} = 10\,020.057, \quad GCV_1 = 10\,024.823, \quad AIC_1 = 9.212123, \quad RMSE_1 = 97.50420;$$

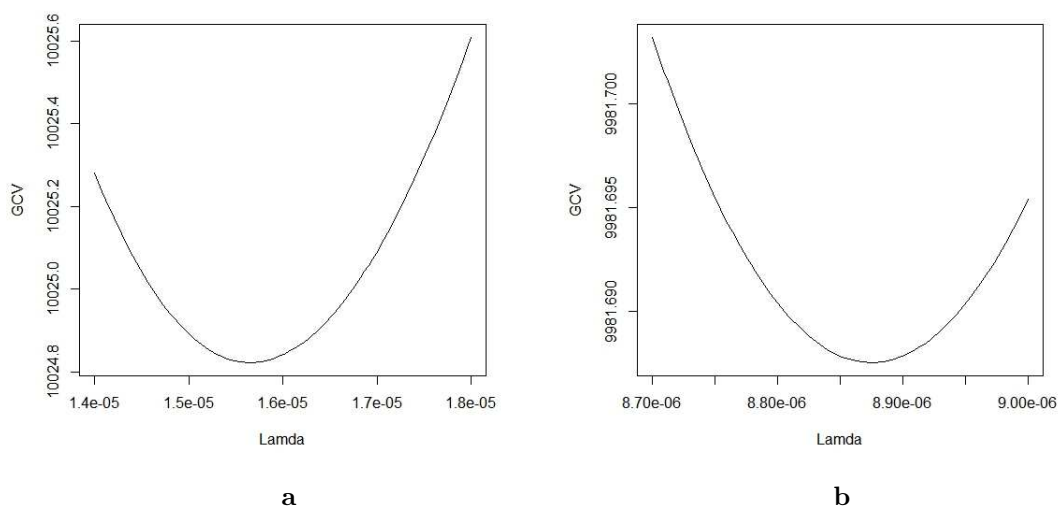


图4 惩罚参数 $\lambda$ 与 $GCV(\lambda)$ 函数关系图(a. 整体惩罚, b. 极差调节的局部惩罚)

图4b采用局部惩罚样条回归, 最优惩罚参数 $\lambda = 8.875758 \times 10^{-6}$ , 各项信息准则得分为

$$C_{p2} = 9977.466, \quad GCV_2 = 9981.688, \quad AIC_2 = 9.207894, \quad RMSE_2 = 97.45515.$$

同样可见, 在各项信息准则下, 极差调节的局部惩罚样条回归模型具有更好的表现.

进一步, 我们将例1独立重复试验100次, 在每次试验中记分别记

$$\begin{aligned} D_1 &= C_{p1} - C_{p2}, & D_2 &= GCV_1 - GCV_2, \\ D_3 &= AIC_1 - AIC_2, & D_4 &= RMSE_1 - RMSE_2. \end{aligned}$$

图5给出了四组数据集 $D_1, D_2, D_3, D_4$ 的箱式图. 由图5可知, 基于极差调节的局部惩罚样条回归模型在整体上比整体惩罚样条回归模型有显著的得分优势. 对例2, 独立重复试验结果显示具有同样的结论.

在上两例中, 我们还将本文所提出的基于极差调节的局部惩罚样条方法与Ruppert和Carroll<sup>[10]</sup>所给的局部惩罚样条回归方法作了试验对比, 模拟结果显示两种方法信息准则得分没有显著差别, 因此该两类方法的模型拟合效果差异很小. 但显然本文所给方法无论在直观意义上, 还是在计算复杂度上都具有明显的优势.

以上结果显示了本文所提出的局部惩罚法相比于整体惩罚法在拟合随机数据上的优越性. 为了进一步检验本文所给出的方法在拟合真实函数上的精确性, 我们在100次模拟实验中, 记录了如下两个指标, 分别是拟合值与真实函数值的平均偏差平方和

$$MSB = n^{-1} \sum_{i=1}^n (f(x_i) - \hat{f}(x_i))^2$$

以及偏差平方和平方根

$$RSB = \sqrt{\sum_{i=1}^n (f(x_i) - \hat{f}(x_i))^2}.$$



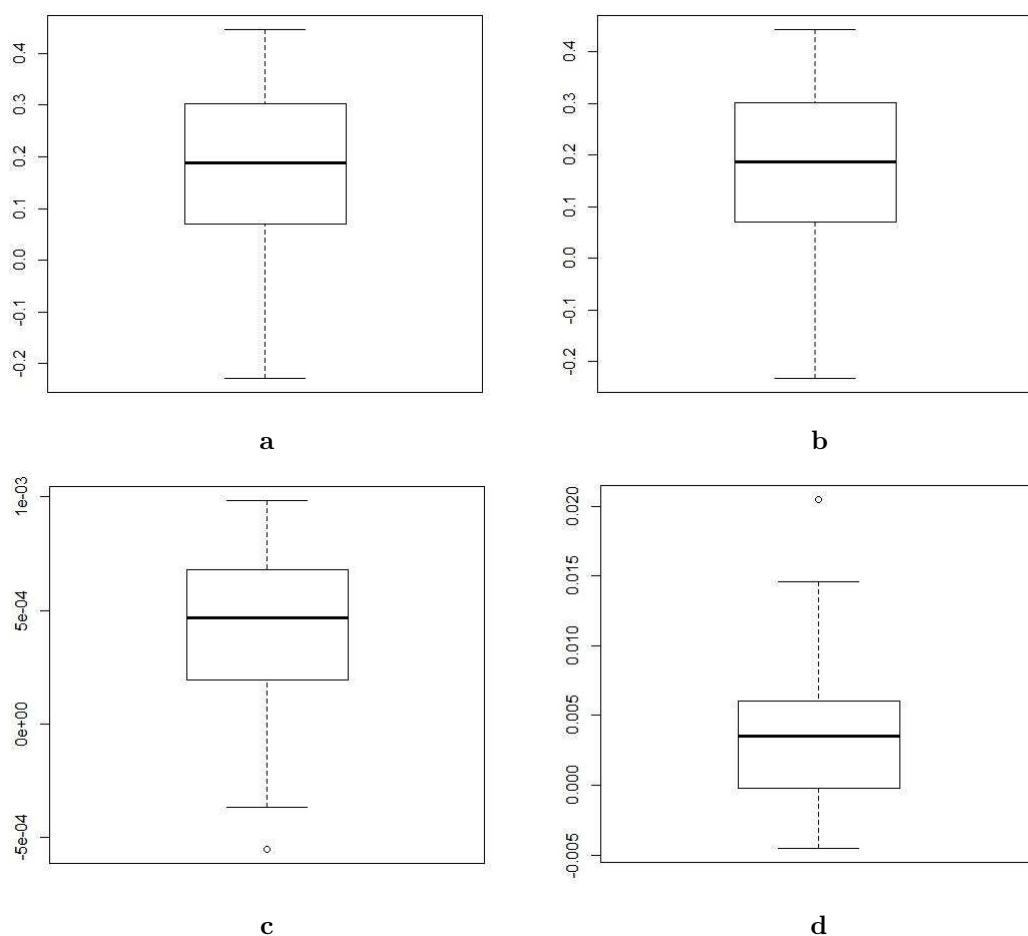


图5 100次独立重复试验下, a, b, c, d依次对应于 $D_1, D_2, D_3, D_4$ 的值

在例1中记

$$D_1 = \text{MSB}_1 - \text{MSB}_2, \quad D_2 = \text{RSB}_1 - \text{RSB}_2,$$

在例2中记

$$D_3 = \text{MSB}_1 - \text{MSB}_2, \quad D_4 = \text{RSB}_1 - \text{RSB}_2.$$

图6给出了 $D_1, D_2, D_3, D_4$ 的箱式图. 由图6可知, 局部惩罚模型比整体惩罚模型在拟合真实函数时偏差较小, 拟合结果更具精确性.

## §5. 总 结

当数据具有空间异质性时, 构造由数据驱动的局部惩罚权重, 更能提升惩罚样条回归模型的总体效果. 本文基于样条截断幂基函数系数的直观意义, 指出当数据在局部范围内具有较大的波动性时, 也应允许此处的拟合曲线具有较大的波动性, 因此应给予较小的惩

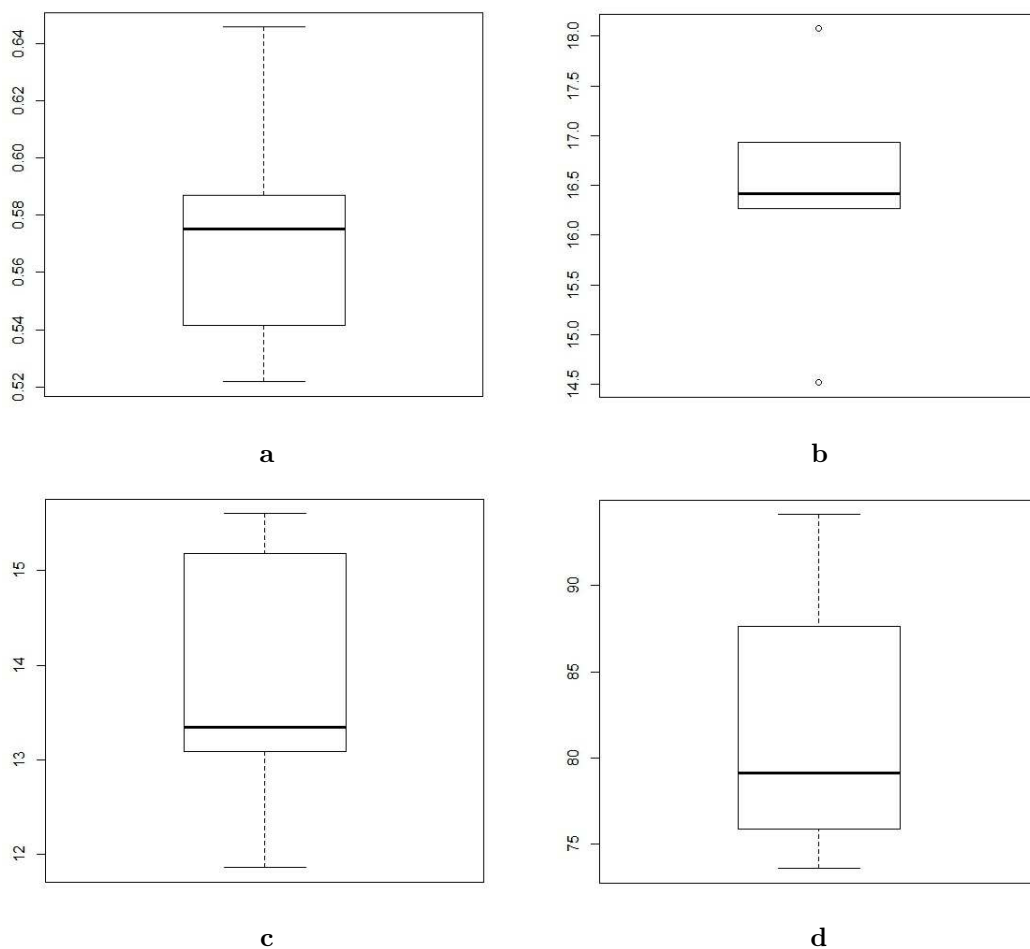


图6 100次独立重复试验下, a, b, c, d依次对应于 $D_1, D_2, D_3, D_4$ 的值

罚. 在此意义上, 本文构造了以极差的线性单减函数作为局部惩罚权重的方法. 模拟结果显示, 采用该方法的惩罚样条回归模型相比整体惩罚模型, 无论是对随机数据的拟合优度, 还是对真实函数的拟合精确度都有更好的信息准则得分. 且相比Ruppert和Carroll<sup>[10]</sup>提出的局部惩罚样条回归模型, 该方法虽然在信息准则得分上几乎没有差别, 但在直观意义和计算复杂度上具有明显的优势.

## 参 考 文 献

- [1] 陈希孺, 倪国熙. 数理统计学教程[M]. 合肥: 中国科学技术大学出版社, 2009.
- [2] 何晓群, 刘文卿. 应用回归分析[M]. 第3版. 北京: 中国人民大学出版社, 2011.
- [3] Green P J, Silverman B W. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*[M]. London: Chapman and Hall/CRC, 1993.
- [4] Eubank R L. *Spline Smoothing and Nonparametric Regression*[M]. New York: Marcel Dekker, 1988.



- [5] 高采文, 甘华来. 增长曲线模型的非参数估计 [J]. 应用概率统计, 2013, **29(6)**: 655–665.
- [6] 王江峰, 梁汉营, 范国良. 左截断相依数据下非参数回归的局部M估计 [J]. 中国科学: 数学, 2012, **42(10)**: 995–1015.
- [7] Gu C. *Smoothing Spline ANOVA Models* [M]. New York: Springer-Verlag, 2002.
- [8] Eilers P H C, Marx B D. Flexible smoothing with *B*-splines and penalties [J]. *Statist. Sci.*, 1996, **11(2)**: 89–102.
- [9] Ruppert D, Wand M P, Carroll R J. *Semiparametric Regression* [M]. New York: Cambridge University Press, 2003.
- [10] Ruppert D, Carroll R J. Spatially-adaptive penalties for spline fitting [J]. *Aust. N. Z. J. Stat.*, 2000, **42(2)**: 205–223.

## Local Penalized Spline Regression Model Based on Range

LI Meng    YANG Lianqiang    JIANG Kun    HE Shengxian

(School of Mathematical Sciences, Anhui University, Hefei, 230601, China)

**Abstract:** Inspired by intuitive meanings of truncated power basis's coefficients, the local penalization based on range's linear decreasing function is given in penalized spline regression model. This method gives less penalization to fitting curve where data is with more volatility, which makes fitted curve controls tradeoff between goodness-of-fit and smoothness better. Simulations show that regression models with local penalized spline obtain lower information rules' scores than global penalized spline when the data is with heteroskedasticity.

**Keywords:** global penalized spline; local panelized spline; heteroskedasticity; range

**2010 Mathematics Subject Classification:** 62G08