一类函数型回归模型的局部扰动分析

周杰*, 卫泽林, 胡黎明

西安电子科技大学数学与统计学院, 西安 710071

E-mail: zhoujie@xidian.edu.cn, 1152050291@qq.com, lm_sunshine6@163.com

收稿日期: 2017-12-30; 接受日期: 2018-04-15; 网络出版日期: 2019-02-26; * 通信作者

国家自然科学基金 (批准号: 71271165) 资助项目

摘要 本文对一类特殊的函数型回归模型, 提出连续型局部扰动分析的概念; 对自变量、因变量和权重函数等三种情形下的扰动分析, 基于泛函优化理论, 推导出扰动函数的表达式. 作为应用, 本文考虑常微分方程的统计诊断问题, 发现了常微分方程两步估计的边界效应, 从而为两步估计中权重函数的选取提供了理论基础; 利用模拟和实际数据验证了统计诊断的有效性.

关键词 函数型回归模型 冗余参数 局部扰动分析 诊断函数 边界效应

MSC (2010) 主题分类 62G10, 62H11

1 引言

函数型数据分析研究当数据是函数时的统计推断问题, 自 Ramsay^[1] 的开创性工作以来, 相关研究工作已得到了长足的发展, 关于函数型主成分分析的研究, 参见文献 [2]; 关于函数型数据的回归分析, 参见文献 [3]; 关于函数型数据的时间序列分析, 参见文献 [4]; 关于函数型比对分析, 参见文献 [5, 6].

与经典统计类似, 函数型数据分析结果也会受到异常观测的影响, 在这种情形下, 采用稳健的统计方法是可行的. Boente 和 Salibián-Barrera^[7] 研究了函数型主成分分析的稳健估计, 利用 Davies^[8] 提出的 S 估计的概念, 定义了尺度参数的稳健估计, 在此基础上, 得到了函数型主成分的稳健估计. Bali 等^[9] 则采用投影寻踪 (projection-pursuit) 方法, 先利用稳健方法把函数型数据投影到单位球面上, 然后研究投影数据的主成分分析, 并在给定的条件下, 得到了估计的相合性. Crambes 等^[10] 则研究了函数型回归模型的稳健估计, 针对因变量是常数, 自变量是曲线的回归模型, 考虑了参数函数的非参数稳健估计, 在独立和 α -混合两种情形下, 证明了估计的矩收敛和 L_p 收敛性. 处理异常观测的另外一种方法是先找到异常观测, 然后分析异常观测的原因, 针对具体情况采用不同的处理方法. 实际上, 文献 [7] 在研究稳健主成分估计时也提出了两种探测异常观测的方法.

本文针对一类特殊的函数型回归模型, 提出一种新的统计诊断方法. 具体地, 我们考虑了因变量是函数、自变量也是函数的回归模型. 这种回归模型包括了常用的常微分方程模型. 为实现统计诊断

英文引用格式: Zhou J, Wei Z L, Hu L M. Local perturbation analysis of functional regression with applications (in Chinese). Sci Sin Math, 2019, 49: 717–730, doi: 10.1360/N012017-00047

的目标, 我们提出了连续型局部扰动分析的概念. 这种方法可以看作通常局部扰动分析在函数型回归模型中的推广. 在存在与不存在冗余参数的两种情形下, 通过求解给定的泛函优化问题, 我们得到了所谓诊断函数的具体表达式. 通过诊断函数可以发现参数估计的强影响区域, 这为改进参数估计提供了可能. 特别地, 利用上述方法, 我们发现了常微分方程两步估计的边界效应, 与已有的研究相吻合. 需要说明的是, 利用连续函数进行扰动分析在统计学中早有先例. 在稳健统计领域, 著名的影响函数、崩溃点都可以看作是所谓极大偏差 (maximum bias) 的特例 (参见文献 [3, 11, 12]). 对给定的统计量和分布模型, 极大偏差对分布进行扰动分析, 其中扰动函数即为一连续函数, 在所有的扰动函数中寻找使得统计量变化最大的扰动. 本文把这种思想应用到一类函数型回归模型中, 得到的结果与常微分方程统计推断已有研究相互验证, 也可用于更一般的模型诊断.

本文结构如下: 第 2 节给出主要理论结果; 第 3 节作为应用, 对线性常微分方程模型给出相关公式; 第 4 节给出模拟数据分析; 第 5 节考虑了一个实际数据的例子; 第 6 节则给出简单的总结和展望.

2 函数型回归模型的局部扰动分析

考虑以下函数型多元非线性回归模型:

$$Y(t) = F(X(t), \theta) + \varepsilon(t), \quad t \in [0, 1], \quad (2.1)$$

其中 $X(t) = (X_1(t), \dots, X_m(t))^T$ 为 m 维自变量; $Y(t) = (Y_1(t), \dots, Y_m(t))^T$ 为因变量; $\theta = (\theta_1, \dots, \theta_d)^T$ 为未知参数; $F = (F_1, \dots, F_m)^T$ 为 m 个回归函数; $\varepsilon(t) = (\epsilon_1(t), \epsilon_2(t), \dots, \epsilon_m(t))^T$ 为误差项, 满足 $E\varepsilon_i(t) = 0$, $\text{Cov}(\varepsilon_i(s), \varepsilon_j(t)) = \Sigma(s, t)$, $\forall 1 \leq i, j \leq m$. 这里假定自变量和因变量具有相同的维数 m , 一方面是为了简化符号表示, 另一方面是因为在常微分方程这种重要情形, 这种假设是成立的. 自变量 $X(t)$ 和因变量 $Y(t)$ 通常通过非参数方法, 如核光滑、样条光滑等方法从离散观测点估计出来, 这里假设这些估计值分别表示为 $a(t) = (a_1(t), \dots, a_m(t))^T$ 和 $b(t) = (b_1(t), \dots, b_m(t))^T$, 则参数 θ 可通过求解以下最优化问题得出:

$$\hat{\theta} = \arg \min_{\theta} \int_0^1 \|b(t) - F(a(t), \theta)\|_2^2 w(t) dt = \arg \min_{\theta} M(\theta), \quad (2.2)$$

其中 $w(t)$ 为权重函数, $\|\cdot\|$ 为 m 维的 Euclid 空间范数. 在一定的条件下, $\hat{\theta}$ 的相合性与 $w(t)$ 无关, 但权重函数的选取可以影响 $\hat{\theta}$ 的渐近方差. 在理想情况下, $w(t)$ 的选取应该反映出 $a(t)$ 、 $b(t)$ 的偏差和方差及 $\varepsilon(t)$ 的方差, 但在实际中这通常是不可行的. 实践中最常用的是取 $w(t) = 1$. 考虑到非参数回归的边界效应, Gugushvili 和 Klaassen^[13] 也提出了一些对边界进行修正的权重函数. 但对给定的问题, 如何选择和评估给定的权重函数还是一个困难的问题. 下面将从自变量估计 $a(t)$ 、因变量估计 $b(t)$ 和权重函数 $w(t)$ 等三个方面来直接评估它们对于参数估计 $\hat{\theta}$ 的影响, 利用我们所得到的诊断函数, 可以直接判断这些函数的各个部分对于参数估计影响的大小, 从而为改进统计推断—例如, 选取权重函数—提供可行基础. 我们将根据模型是否存在冗余参数, 分别讨论相关的统计诊断步骤.

2.1 无冗余参数时的扰动分析

考虑如何局部改变模型 (2.1) 和 (2.2) 能够对参数估计 $\hat{\theta}$ 产生最大影响, 这里局部指的是在给定的 $a(t)$ 、 $b(t)$ 和 $w(t)$ 的小邻域内. 首先考虑 $a(t)$ 的扰动分析. 令 $\tilde{a}(t)$ 表示 $a(t)$ 的第 i 个元素受到

扰动后的值, 即 $\tilde{a}(t) = a + e_i \varepsilon p(t) = (a_1(t), \dots, a_{i-1}(t), a_i(t) + \varepsilon p(t), a_{i+1}(t), \dots, a_m(t))$, 其中 $\varepsilon > 0$, $p(t) \in C([0, 1])$ 且满足条件 $\|p(t)\|_{L^2([0, 1])}^2 = \int_0^1 p^2(t) dt = 1$. 类比 (2.2), 定义扰动后参数 θ 的估计为

$$\hat{\theta}(p, \varepsilon) = \arg \min_{\theta} \int_0^1 \|b(t) - F(\tilde{a}(t), \theta)\|_2^2 w(t) dt = \arg \min_{\theta} M(\theta | p, \varepsilon), \quad (2.3)$$

其中

$$M(\theta | p, \varepsilon) = \int_0^1 \|b(t) - F(\tilde{a}(t), \theta)\|_2^2 w(t) dt. \quad (2.4)$$

注意到, 对于任意给定的 $p(t)$, $\hat{\theta} = \hat{\theta}(p, 0)$. 下面考虑 $\hat{\theta}(p, \varepsilon)$ 沿方向 $p(t)$ 在 $a(t)$ 处的 Gâteaux 导数, 定义

$$\lim_{\varepsilon \rightarrow 0} \frac{\hat{\theta}(p, \varepsilon) - \hat{\theta}}{\varepsilon} = \lim_{\varepsilon \rightarrow 0} \frac{\hat{\theta}(p, \varepsilon) - \hat{\theta}(p, 0)}{\varepsilon} \triangleq l(p). \quad (2.5)$$

作为导数, $l(p)$ 衡量了参数估计在 $p(t)$ 方向上的变化速度. 由于 $l(p)$ 是向量, 故定义以下广义 Cook 距离 GCD (generalized Cook distance):

$$\text{GCD}(p) = l^T(p) \Omega l(p), \quad (2.6)$$

其中 Ω 是非负定的权重矩阵, 常用的如单位阵、 $\hat{\theta}$ 的协方差矩阵等. 如果广义 Cook 距离 GCD 在给定的扰动函数 $p(t)$ 达到最大值, 则称此 $p(t)$ 为诊断函数, 在 $p(t)$ 方向的微小改变可以对参数估计产生最大的影响, 据此我们可以对 $a(t)$ 进行统计诊断. 我们需求解以下约束优化问题:

$$\max_{\|p(t)\|_2=1} \text{GCD}(p). \quad (2.7)$$

这里需要对 $p(t)$ 加上约束条件 $\|p(t)\|_2 = 1$, 如果没有上述约束, $\text{GCD}(p)$ 的值将不存在最大值.

对优化问题 (2.3), 由 $\hat{\theta}(p, \varepsilon)$ 的定义, 可得

$$\left. \frac{\partial M(\theta | p, \varepsilon)}{\partial \theta} \right|_{\theta=\hat{\theta}(p, \varepsilon)} = 0, \quad (2.8)$$

由 (2.8) 和 Taylor 展开公式, 可得

$$\left. \frac{\partial M(\theta | p, \varepsilon)}{\partial \theta} \right|_{\theta=\hat{\theta}} + \left. \frac{\partial M(\theta | p, \varepsilon)}{\partial \theta \partial \theta^T} \right|_{\theta=\hat{\theta}} (\hat{\theta}(p, \varepsilon) - \hat{\theta}) + O_p(\varepsilon^2) = 0, \quad (2.9)$$

也即

$$\hat{\theta}(p, \varepsilon) - \hat{\theta} = \left[\frac{\partial^2 M(\theta | p, \varepsilon)}{\partial \theta \partial \theta^T} \right]^{-1} \left. \frac{\partial M(\theta | p, \varepsilon)}{\partial \theta} \right|_{\theta=\hat{\theta}} + O_p(\varepsilon^2). \quad (2.10)$$

由 $M(\theta | p, \varepsilon)$ 的定义, 对一阶导数, 有

$$\begin{aligned} \left. \frac{\partial M(\theta | p, \varepsilon)}{\partial \theta} \right|_{\theta=\hat{\theta}} &= -2 \int_0^1 \dot{F}_2^T(\tilde{a}, \theta) (b(t) - F(\tilde{a}, \theta)) w(t) dt \Big|_{\theta=\hat{\theta}} \\ &= -2\varepsilon \int_0^1 [\ddot{F}_{21}^T(b - F) - \dot{F}_2^T \dot{F}_1] w(t) p dt + O(\varepsilon^2), \end{aligned} \quad (2.11)$$

其中

$$\ddot{F}_{21}^T(b - F) = \sum_{j=1}^m \frac{\partial^2 F_j}{\partial \theta \partial a_i} (b_j - F_j), \quad \dot{F}_2^T \dot{F}_1 = \sum_{j=1}^m \frac{\partial F_j}{\partial \theta} \frac{\partial F_j}{\partial a_i},$$

上述函数都是在 $(a(t), \hat{\theta})$ 处赋值. 这里利用了以下事实: $\int_0^1 \dot{F}_2^T(a, \hat{\theta})(b - F(a, \hat{\theta}))w(t)dt = 0$. 对二阶导数, 有

$$\left. \frac{\partial^2 M(\theta | p, \epsilon)}{\partial \theta \partial \theta^T} \right|_{\theta=\hat{\theta}} = -2 \int_0^1 [\ddot{F}_{22}^T(b - F) - \dot{F}_2^T \dot{F}_2]w(t)dt + O(\epsilon) \triangleq 2A + O(\epsilon), \quad (2.12)$$

其中 $\ddot{F}_{22}^T(b - F) = \sum_{i=1}^m \frac{\partial^2 F_i}{\partial \theta \partial \theta^T}(b_i - F_i)$, $\dot{F}_2^T \dot{F}_2 = \sum_{i=1}^m \frac{\partial F_i}{\partial \theta} \frac{\partial F_i}{\partial \theta^T}$. 函数 F , $\dot{F}_2$ 和 $\ddot{F}_{22}$ 都在 $(a(t), \hat{\theta})$ 处赋值. 基于 (2.11) 和 (2.12), 有

$$\lim_{\epsilon \rightarrow 0} \frac{\hat{\theta}(p, \epsilon) - \hat{\theta}}{\epsilon} = A^{-1}(0) \int_0^1 [\ddot{F}_{21}^T(b - F) - \dot{F}_2^T \dot{F}_1]w(t)p dt = A^{-1}(0) \int_0^1 g(t)p dt, \quad (2.13)$$

其中

$$g(t) = w(t)(g_1(t), \dots, g_d(t))^T = w(t)(\ddot{F}_{21}^T(b - F) - \dot{F}_2^T \dot{F}_1). \quad (2.14)$$

将 (2.13) 代入 (2.6) 可得

$$\text{GCD}(p) = \left\{ \int_0^1 g^T(t)p(t)dt \right\} A^{-1} \Omega A^{-1} \left\{ \int_0^1 g(t)p(t)dt \right\}. \quad (2.15)$$

为求解约束优化问题 (2.7), 引入以下 Lagrange 泛函:

$$L(p, \lambda) = \text{GCD}(p) + \lambda \left(\int_0^1 p^2(t)dt - 1 \right). \quad (2.16)$$

记 $DL(p(t), \lambda)h(t)$ 为 $L(p, \lambda)$ 在 $p(t)$ 处沿任意给定方向 $h(t) \in C([0, 1])$ 的 Gâteaux 导数, 则由泛函优化的必要条件知, 若 $L(p, \lambda)$ 在 $p(t)$ 处达到最大值, 则必有

$$DL(p(t), \lambda)h(t) = \lim_{\epsilon \rightarrow 0} \frac{L(p(t) + \epsilon h(t), \lambda) - L(p(t), \lambda)}{\epsilon} = 0. \quad (2.17)$$

因此, 将 (2.16) 代入上式得

$$\int_0^1 \left\{ \left[\int_0^1 g(t)p(t)dt \right] A^{-1} \Omega A^{-1} g(t) - \lambda p(t) \right\} h(t)dt = 0. \quad (2.18)$$

因为以上等式对空间 $C([0, 1])$ 上所有的函数 $h(t)$ 都成立, 故有

$$\left[\int_0^1 g(t)p(t)dt \right] A^{-1} \Omega A^{-1} g(t) = \lambda p(t). \quad (2.19)$$

以上等式意味着存在一个向量 $\alpha = (\alpha_1, \dots, \alpha_d)^T$ 满足 $p(t) = \alpha^T g(t)$. 将 $p(t) = \alpha^T g(t)$ 代回等式 (2.19), 可得

$$\alpha^T \left[\int_0^1 g(t)g^T(t)dt \right] A^{-1} \Omega A^{-1} g(t) = \lambda \alpha^T g(t). \quad (2.20)$$

不失一般性, 假设 $g(t) = (g_1(t), \dots, g_d(t))^T$ 各元素之间线性不相关, 则有

$$\left[A^{-1} \Omega A^{-1} \int_0^1 g(t)g^T(t)dt \right] \alpha = \lambda \alpha, \quad (2.21)$$

即系数 α 为矩阵 $A^{-1}\Omega A^{-1} \int_0^1 g(t)g^T(t)dt$ 特征值 λ 所对应的特征向量. 将 $p(t) = \alpha^T g(t)$ 代入到 (2.15), 则有

$$\begin{aligned} \text{GCD}(p) &= \alpha^T \int_0^1 g(t)g^T(t)dt \left[A^{-1}\Omega A^{-1} \int_0^1 g(t)g^T(t)dt \right] \alpha \\ &= \lambda \alpha^T \left[\int_0^1 g(t)g^T(t)dt \right] \alpha = \lambda, \end{aligned} \quad (2.22)$$

其中第二个等式利用了约束条件 $\|p(t)\|_2 = 1$. 由此可知, 诊断函数 $p(t)$ 可以表示为 $p(t) = \alpha^T g(t)$, 其中 α 为矩阵

$$\Phi = A^{-1}\Omega A^{-1} \int_0^1 g(t)g^T(t)dt \quad (2.23)$$

的最大特征值 $\lambda_{\max}$ 所对应的特征向量, $\lambda_{\max}$ 为 GCD 的局部最大值, 而 $g(t)$ 由 (2.14) 给出.

下面考虑 $b(t)$ 的扰动分析. 与 $a(t)$ 下的情形类似, 定义

$$\tilde{b}(t) = b(t) + e_i \varepsilon p(t) = (b_1(t), \dots, b_{i-1}(t), b_i(t) + \varepsilon p(t), b_{i+1}(t), \dots, b_m(t)). \quad (2.24)$$

有了 $\tilde{b}(t)$, 则扰动后参数 θ 的估计通过求解以下优化问题得出:

$$\begin{aligned} \hat{\theta}(p, \varepsilon) &= \arg \min_{\theta} \int_0^1 \|\tilde{b}(t) - F(a(t), \theta)\|^2 w(t) dt \\ &= \arg \min_{\theta} M(\theta | p, \varepsilon). \end{aligned} \quad (2.25)$$

有了 $\hat{\theta}(p, \varepsilon)$, 我们就可以定义如等式 (2.13) 和 (2.6) 所示的 Gâteaux 导数 $l(p)$ 以及广义 Cook 距离 $\text{GCD}(p)$. 采取与对 $a(t)$ 加扰动后求解最优化问题相同的方法, 我们得到与等式 (2.16) 相同的形式, 于是, 诊断函数可以表示为 $p(t) = \alpha^T g(t)$, 其中

$$g(t) = w(t) \frac{\partial F_i}{\partial \theta}, \quad (2.26)$$

α 满足等式 (2.21), 这里 $g(t)$ 为等式 (2.26) 所示, 矩阵 A 与 (2.12) 所定义的相同.

对于权重函数 $w(t)$ 所加扰动的情形, 定义添加扰动后的权重函数为 $\tilde{w}(t) = w(t) + \varepsilon p(t)$. 于是, 参数 θ 的估计由以下最优化问题得出:

$$\begin{aligned} \hat{\theta}(p, \varepsilon) &= \arg \min_{\theta} \int_0^1 \|b(t) - F(a(t), \theta)\|^2 \tilde{w}(t) dt \\ &= \arg \min_{\theta} M(\theta | p, \varepsilon). \end{aligned} \quad (2.27)$$

同样定义 $l(p)$ 和 $\text{GCD}(p)$, 并且求解优化问题 (2.7). 相同的方法可以求得诊断函数 $p(t)$ 的表达式为 $p(t) = \alpha^T g(t)$, 此时,

$$g(t) = w(t) \dot{F}_2^T(b - F), \quad (2.28)$$

系数 α 满足 (2.21), 而矩阵 A 由 (2.12) 所定义. 我们把这些结果总结成以下定理 1.

定理 1 对于函数型多元非线性回归模型 (2.1) 的两步估计 (2.2), 分别考虑三种情形扰动的估计 (2.3)、(2.25) 和 (2.27), 相应的诊断函数都可表示为 $p(t) = \alpha^T g(t)$, 其中三种扰动的 $g(t)$ 分别由 (2.14)、(2.26) 和 (2.28) 定义. 系数 α 是矩阵 $\Phi = A^{-1}\Omega A^{-1} \int_0^1 g(t)g^T(t)dt$ 的最大特征值 $\lambda_{\max}$ 所对应的特征向量, 矩阵 A 由 (2.12) 所定义.

2.2 存在冗余参数时的扰动分析

给定参数 θ , 常常要考虑其中部分参数的统计诊断问题, 这时 θ 可表示为 $\theta = (\theta_1^T, \theta_2^T)^T$, 其中 θ_1 为兴趣参数, θ_2 为冗余参数. 本节讨论参数 θ_1 的统计诊断问题. 不失一般性, 假设 θ_1 为 d_1 维向量, θ_2 为 d_2 维的向量. 对于任意给定的 θ_1 , 定义冗余参数 θ_2 的估计式为

$$\begin{aligned}\tilde{\theta}_2(\theta_1) &= \arg \min_{\theta_2} \int_0^1 \|b(t) - F(a(t), \theta_1, \theta_2)\|^2 w(t) dt \\ &= \arg \min_{\theta_2} M(\theta).\end{aligned}\quad (2.29)$$

有了 θ_2 , 则 θ_1 的估计值通过解决以下最优化问题得出:

$$\begin{aligned}\hat{\theta}_1 &= \arg \min_{\theta_1} \int_0^1 \|b(t) - F(a(t), \theta_1, \tilde{\theta}_2(\theta_1))\|^2 w(t) dt \\ &= \arg \min_{\theta_1} \tilde{M}(\theta_1),\end{aligned}\quad (2.30)$$

显然, $\tilde{\theta}_2(\hat{\theta}_1) = \hat{\theta}_2$.

首先考虑对 $a(t)$ 添加扰动的情形. 设 $\tilde{a}(t) = a(t) + e_i \varepsilon p(t) = (a_1(t), \dots, a_i(t) + \varepsilon p(t), \dots, a_m(t))^T$, 则受到扰动后参数 θ_1 的估计定义为

$$\begin{aligned}\hat{\theta}_1(p(t), \varepsilon) &= \arg \min_{\theta_1} \int_0^1 \|b(t) - F(\tilde{a}(t), \theta_1, \tilde{\theta}_2(\theta_1))\|^2 w(t) dt \\ &= \arg \min_{\theta_1} \tilde{M}(\theta_1 | p(t), \varepsilon),\end{aligned}\quad (2.31)$$

其中

$$\tilde{M}(\theta_1 | p(t), \varepsilon) = \int_0^1 \|b(t) - F(\tilde{a}(t), \theta_1, \tilde{\theta}_2(\theta_1))\|^2 w(t) dt. \quad (2.32)$$

同 (2.9) 类似, 有

$$\frac{\partial \tilde{M}(\theta_1 | p, \varepsilon)}{\partial \theta_1} + \frac{\partial^2 \tilde{M}(\theta_1 | p, \varepsilon)}{\partial \theta_1 \partial \theta_1^T} (\hat{\theta}_1(p, \varepsilon) - \hat{\theta}_1) + O_p(\varepsilon^2) = 0. \quad (2.33)$$

对于上式左边第一项, 计算得

$$\begin{aligned}\frac{\partial \tilde{M}(\theta_1 | p, \varepsilon)}{\partial \theta_1} &= -2 \int_0^1 \left(\dot{F}_2(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) + \dot{F}_3(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T (b - F(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2)) w(t) dt \\ &= -2\varepsilon \int_0^1 \left[\left(\ddot{F}_{21} + \ddot{F}_{31} \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T (b - F) - \left(\dot{F}_2 + \dot{F}_3 \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T \dot{F}_1 \right] p(t) w(t) dt + O(\varepsilon^2),\end{aligned}\quad (2.34)$$

其中 $\dot{F}_2 = \frac{\partial F}{\partial \theta_1}$, $\dot{F}_3 = \frac{\partial F}{\partial \theta_2}$, $\ddot{F}_{21} = \frac{\partial^2 F}{\partial \theta_1 \partial a_i}$, $\ddot{F}_{31} = \frac{\partial^2 F}{\partial \theta_2 \partial a_i}$, 并且所有函数都是在 $(a(t), \hat{\theta}_1, \hat{\theta}_2)$ 处赋值.

对于 $\tilde{M}(\theta_1 | p(t), \varepsilon)$ 的二阶导数, 计算可得

$$\begin{aligned}\frac{\partial^2 \tilde{M}(\theta_1 | p, \varepsilon)}{\partial \theta_1 \partial \theta_1^T} &= -2 \int_0^1 \left[\left(\ddot{F}_{22}(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) + \ddot{F}_{23}(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) \frac{d\tilde{\theta}_2}{d\theta_1^T} \right. \right. \\ &\quad \left. \left. + \left(\ddot{F}_{32}(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) + \ddot{F}_{33}(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) \frac{d\tilde{\theta}_2}{d\theta_1^T} \right) \frac{d\tilde{\theta}_2}{d\theta_1^T} + \dot{F}_3(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) \frac{d^2 \tilde{\theta}_2}{d\theta_1^2} \right)^T (b - F(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2)) \right. \\ &\quad \left. - \left(\dot{F}_2(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) + \dot{F}_3(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T \left(\dot{F}_2(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) + \dot{F}_3(\tilde{a}, \hat{\theta}_1, \hat{\theta}_2) \frac{d\tilde{\theta}_2}{d\theta_1^T} \right) \right] w(t) dt\end{aligned}$$

$$\begin{aligned}
&= -2 \int_0^1 \left[\left(\ddot{F}_{22} + \ddot{F}_{23} \frac{d\tilde{\theta}_2}{d\theta_1} + \left(\ddot{F}_{32} + \ddot{F}_{33} \frac{d\tilde{\theta}_2}{d\theta_1^T} \right) \frac{d\tilde{\theta}_2}{d\theta_1^T} + \dot{F}_3 \frac{d^2\tilde{\theta}_2}{d\theta_1^2} \right)^T (b - F) \right. \\
&\quad \left. - \left(\dot{F}_2 + \dot{F}_3 \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T \left(\dot{F}_2 + \dot{F}_3 \frac{d\tilde{\theta}_2}{d\theta_1^T} \right) \right] w(t) dt + O(\varepsilon^2) \\
&\triangleq 2A + O(\varepsilon),
\end{aligned} \tag{2.35}$$

其中

$$\begin{aligned}
\ddot{F}_{22}^T (b - F) &= \sum_{i=1}^m \frac{\partial^2 F_i}{\partial \theta_1 \partial \theta_1^T} (b_i - F_i), \\
\left(\ddot{F}_{23} \frac{d\tilde{\theta}_2}{d\theta_1} \right)^T (b - F) &= \sum_{i=1}^m \frac{\partial^2 F_i}{\partial \theta_1 \partial \theta_2^T} \frac{d\tilde{\theta}_2}{d\theta_1} (b_i - F_i), \\
\left(\ddot{F}_{33} \frac{d\tilde{\theta}_2}{d\theta_1^T} \left(\frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T \right)^T (b - F) &= \sum_{i=1}^m \sum_{j=1}^{d_2} \sum_{k=1}^{d_2} \frac{\partial^2 F_i}{\partial \tilde{\theta}_{2j} \partial \tilde{\theta}_{2k}} \frac{\partial \tilde{\theta}_{2j}}{\partial \theta_1} \frac{\partial \tilde{\theta}_{2k}}{\partial \theta_1^T} (b_i - F_i), \\
\left(\dot{F}_3 \frac{d^2\tilde{\theta}_2}{d(\theta_1^T)^2} \right)^T (b - F) &= \sum_{i=1}^m \sum_{j=1}^{d_2} \frac{\partial F_i}{\partial \tilde{\theta}_{2j}} \frac{\partial^2 \tilde{\theta}_{2j}}{\partial \theta_1 \partial \theta_1^T} (b_i - F_i).
\end{aligned} \tag{2.36}$$

将等式 (2.34) 和 (2.35) 代入 (2.33), 可得

$$\begin{aligned}
\lim_{\varepsilon \rightarrow 0} \frac{\hat{\theta}_1(p, \varepsilon) - \hat{\theta}_1}{\varepsilon} &= A^{-1} \int_0^1 \left[\left(\ddot{F}_{21} + \ddot{F}_{31} \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T (b - F) - \left(\dot{F}_2 + \dot{F}_3 \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T \dot{F}_1 \right] p(t) w(t) dt \\
&\triangleq l(p).
\end{aligned} \tag{2.37}$$

有了 $l(p)$ 的表达式, 我们可以定义广义 Cook 距离, $\text{GCD}(p) = l^T(p) \Omega l(p)$. 下面要求出诊断函数 $p(t)$ 以最大化 $\text{GCD}(p)$, 其求解方法与上一小节中涉及的方法类似. 经推导可得矩阵 A 的表达式如等式 (2.35) 所示, 而函数 $g(t)$ 的表达式如下所示:

$$g(t) = w(t) \left[\left(\ddot{F}_{21} + \ddot{F}_{31} \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T (b - F) - \left(\dot{F}_2 + \dot{F}_3 \frac{d\tilde{\theta}_2}{d\theta_1^T} \right)^T \dot{F}_1 \right]. \tag{2.38}$$

求得 A 和 $g(t)$ 之后, 即可得到诊断函数 $p(t) = \alpha^T g(t)$, 其中 α 为矩阵 $\Phi = A^{-1} \Omega A^{-1} \int_0^1 g(t) g^T(t) dt$ 的最大特征值 $\lambda_{\max}$ 所对应的特征向量.

下面讨论当 $b(t)$ 受到扰动后的情形. 定义 $b(t)$ 受到扰动后的状态为 $\tilde{b}(t) = b + \varepsilon e_i p(t)$, 此时参数估计为以下约束优化问题的求解:

$$\hat{\theta}_1(p(t), \varepsilon) = \arg \min_{\theta_1} \int_0^1 \|\tilde{b}(t) - F(a(t), \theta_1, \tilde{\theta}_2(\theta_1))\|^2 w(t) dt. \tag{2.39}$$

同样定义 $\text{GCD}(p)$, 同样的方法得到诊断函数 $p(t) = \alpha^T g(t)$, 此时, $g(t)$ 的表达式为

$$g(t) = w(t) \left(\dot{F}_2 + \dot{F}_3 \frac{d\tilde{\theta}_2}{d\theta_1} \right)^T dt, \tag{2.40}$$

$\hat{\alpha}$ 为矩阵 $\Phi = A^{-1} \Omega A^{-1} \int_0^1 g(t) g^T(t) dt$ 的最大特征值 $\lambda_{\max}$ 所对应的特征向量. 矩阵 A 由 (2.35) 定义.

接下来讨论权重函数 $w(t)$ 受到扰动后的情形. 假设 $w(t)$ 受到扰动后的状态为 $\tilde{w}(t) = w(t) + \varepsilon p(t)$, 此时参数估计定义为

$$\hat{\theta}_1(p(t), \varepsilon) = \arg \min_{\theta_1} \int_0^1 \|b(t) - F(a(t), \theta_1, \tilde{\theta}_2(\theta_1))\|^2 \tilde{w}(t) dt. \tag{2.41}$$

同样得到诊断函数 $p(t) = \alpha^T g(t)$, 此时, $g(t)$ 的表达式为

$$g(t) = w(t) \left(\dot{F}_2 + \dot{F}_3 \frac{d\tilde{\theta}_2}{d\theta_1} \right)^T (b - F), \quad (2.42)$$

α 为矩阵 $\Phi = A^{-1} \Omega A^{-1} \int_0^1 g(t) g^T(t) dt$ 的最大特征值 $\lambda_{\max}$ 所对应的特征向量, 矩阵 A 同上. 我们把这些结果总结成以下定理以方便研读.

定理 2 对于函数型多元非线性回归模型 (2.1) 的两步估计 (2.2), 假设 $\theta = (\theta_1^T, \theta_2^T)^T$, 其中 θ_1 为兴趣参数, 分别考虑三种情形扰动的估计 (2.31)、(2.39) 和 (2.41), 相应的诊断函数都可表示为 $p(t) = \alpha^T g(t)$, 其中三种扰动的 $g(t)$ 分别由 (2.38)、(2.40) 和 (2.42) 定义. 系数 α 是矩阵

$$\Phi = A^{-1} \Omega A^{-1} \int_0^1 g(t) g^T(t) dt$$

的最大特征值 $\lambda_{\max}$ 所对应的特征向量. 矩阵 A 由 (2.35) 所定义.

3 一个特例: 线性常微分方程组的扰动分析

众所周知, 常微分方程的参数估计由于涉及常微分方程的数值算法, 运算量很大, 是统计学中的难题之一, 特别是对高维常微分方程组. 近年来, 基于非参数回归的两步估计在常微分方程的统计推断领域受到很大的关注, 两步估计不用求解常微分方程, 并且利用两步估计也较容易处理高维的常微分方程组 (参见文献 [13–15]). 两步估计可以表示为模型 (2.1) 的形式, 其中 (2.2) 中的向量 b 表示状态变量导数的估计, a 表示状态变量自身的估计, 而未知参数 θ 的估计即通过优化问题 (2.2) 得到, 所以, 上面所讨论的扰动分析方法特别适用于常微分方程的统计诊断. 下面针对三维的线性常微分方程组, 给出诊断函数 $p(t)$ 的表达式.

考虑线性常微分方程 $\dot{X}(t) = GX + c$, 其中 $X(t) = (X_1(t), X_2(t), X_3(t))^T$, $c = (c_1, c_2, c_3)^T$, G 是系数矩阵, 这里 β_{ij} 表示 G 的第 i 行第 j 列的元素. 与上一节相同, 用 $b(t) = (b_1(t), b_2(t), b_3(t))^T$ 表示导数 $\dot{X}(t) = (\dot{X}_1(t), \dot{X}_2(t), \dot{X}_3(t))^T$ 的估计, 用 $a(t) = (a_1(t), a_2(t), a_3(t))^T$ 表示 $X(t)$ 的估计, 未知参数为 $\theta = (\beta_{11}, \beta_{12}, \beta_{13}, c_1, \beta_{21}, \beta_{22}, \beta_{23}, c_2, \beta_{31}, \beta_{32}, \beta_{33}, c_3)^T$. θ 的估计 $\hat{\theta}$ 由优化问题 (2.2) 得到, 而 $\hat{G}$ 表示矩阵 G 内的元素由估计量替换后的矩阵. 我们仅考虑 $a_1(t)$ 、 $b_1(t)$ 和 $w(t)$ 的扰动分析, 其他关于 $a_2(t)$ 、 $a_3(t)$ 、 $b_2(t)$ 和 $b_3(t)$ 的相关分析可以类似得到.

注意到在 $a_1(t)$ 、 $b_1(t)$ 和 $w(t)$ 的扰动方案下, 影响函数 $p(t)$ 具有相同的形式 $p(t) = \alpha^T g(t)$, 其中由 (2.14), $a_1(t)$ 所对应的 $g(t)$ 为

$$g(t) = w(t) \left((b - \hat{G}a - \hat{c}) \otimes \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \hat{G}_{\cdot 1} \otimes \begin{pmatrix} a(t) \\ 1 \end{pmatrix} \right), \quad (3.1)$$

这里 $\otimes$ 表示矩阵的 Kronecker 乘积, $\hat{G}_{\cdot 1}$ 表示矩阵 $\hat{G}$ 的第一列; 由 (2.26), $b_1(t)$ 所对应的 $g(t)$ 为

$$g(t) = w(t)(a_1(t), a_2(t), a_3(t), 1, 0, 0, 0, 0, 0, 0, 0, 0)^T. \quad (3.2)$$

由 (2.28), $w(t)$ 所对应的 $g(t)$ 为

$$g(t) = (b - \hat{G}a(t) - \hat{c}) \otimes \begin{pmatrix} a \\ 1 \end{pmatrix}. \quad (3.3)$$

对 $a_1(t)$ 、 $b_1(t)$ 和 $w(t)$ 三种扰动方案, 它们对应的系数向量 α 也具有相同的形式, 即都为 (2.23) 中所定义矩阵 Φ 最大特征根所对应的特征向量, 其中矩阵 Φ 的定义式中所涉及的矩阵 A 可以表示为 $A = I_3 \otimes H$, 这里 I_3 是三阶单位阵, 矩阵 H 为

$$H = \int_0^1 \begin{pmatrix} a(t)a^T(t) & a(t) \\ a^T(t) & 1 \end{pmatrix} w(t) dt. \quad (3.4)$$

下面考虑在冗余参数情形下的影响函数. 不失一般性, 假定兴趣参数为 $\theta_1 = (\beta_{11}, \beta_{12}, \beta_{13}, c_1)^T$, 则对 $a_1(t)$ 、 $b_1(t)$ 和 $w(t)$ 三种扰动方案, $p(t)$ 依然具有形式 $p(t) = \alpha^T g(t)$, 其中对于 $a_1(t)$, 由 (2.38) 得

$$g(t) = w(t) \begin{pmatrix} b_1 - \hat{G}_1^T a(t) - \hat{c}_1 - \hat{\beta}_{11} a_1(t) \\ -\hat{\beta}_{11} a_2(t) \\ -\hat{\beta}_{11} a_3(t) \\ -\hat{\beta}_{11} \end{pmatrix}, \quad (3.5)$$

这里 $\hat{G}_1$ 表示 $\hat{G}$ 的第一行; 对于 $b_1(t)$, 由 (2.40) 有

$$g(t) = w(t)(a_1(t), a_2(t), a_3(t), 1)^T; \quad (3.6)$$

对于 $w(t)$, 由 (2.42) 有

$$g(t) = w(t)(b_1(t) - \hat{G}_1^T a(t) - \hat{c}_1)(a_1(t), a_2(t), a_3(t), 1)^T. \quad (3.7)$$

而 α 亦为矩阵 Φ 最大特征根所对应的特征向量, 其中 Φ 中所涉及的矩阵 A 即为 (3.4) 所定义的 H 矩阵.

4 模拟数据分析

本节以 FitzHugh-Nagumo 方程为例, 利用模拟数据研究上述统计诊断方法的特点. FitzHugh-Nagumo 方程由 FitzHugh^[16] 和 Nagumo 等^[17] 建立, 可看作 Hodgkin-Huxley 模型的简化, 用以描述高峰电位在鱿鱼神经元的巨型轴突中的表现, 具体如下:

$$\begin{cases} \dot{V} = c \left(V - \frac{V^3}{3} + R \right), \\ \dot{R} = -\frac{1}{c} (V - a + bR), \end{cases} \quad (4.1)$$

其中 V 表示穿过轴突膜的电压, R 表示汇总外向电流, 未知参数为 $\theta = (a, b, c)^T$. 这里设定参数值分别为 (0.2, 0.2, 3). 利用四阶的 Runge-Kutta 算法在区间 $[0, 20]$ 内, 以 $V_0 = -1$ 和 $R = 1$ 为初始值生成

模型的数值解, 步长为 $h = 0.01$. 然后从其中等间隔选取 40 个点, 对其添加正态噪声得到模拟观测值如下:

$$Y_V(t_i) = V(t_i) + \epsilon_{1i}, \quad Y_R(t_i) = R(t_i) + \epsilon_{2i}, \quad i = 1, \dots, 40, \quad (4.2)$$

其中 $\epsilon_{ij} \sim N(0, 0.04^2)$, $i = 1, 2, j = 1, \dots, 40$. 基于这些观测值和局部核光滑, 得到状态变量及其导数的估计 $a(t)$ 和 $b(t)$, 其中所用到的核函数为正态核函数, 窗宽选取根据均方误差最小的最优窗宽原则选取 (详细参见著作 [18, 第 3 章]).

利用第 2 节的记号, 记 $b_1(t)$ 和 $b_2(t)$ 分别为导数 $\dot{V}(t)$ 和 $\dot{R}(t)$ 的估计, $a_1(t)$ 和 $a_2(t)$ 为状态变量 $V(t)$ 和 $R(t)$ 的估计, 即 $a(t) = (a_1(t), a_2(t))^T$, $b(t) = (b_1(t), b_2(t))^T$. 参数 θ 的估计根据 (2.2) 所示的方法得出. 记 $e = (e_1, e_2)^T = b - F$ 为残差向量. 这里先考虑 $a(t)$ 的扰动分析.

由 (2.12) 可知, 对于以上模型有 $A = H_1 + H_2$, 其中 $H_1 = -\int_0^1 \ddot{F}_{22}^T(b - F)dt$, $H_2 = \int_0^1 \dot{F}_2^T \dot{F}_2 dt$. 对状态变量 $V(t)$ 和 $R(t)$, $g(t)$ 函数表达式分别为

$$\begin{aligned} g_1(t) &= \begin{pmatrix} \frac{1}{c^2} \\ -\frac{1}{c^2}a_2 \\ e_1(1 - a_1^2) + \frac{1}{c^2}e_2 - c(1 - a_1^2)(a_1 - \frac{a_1^3}{3} + a_2) + \frac{1}{c^3}(a_1 - a + ba_2) \end{pmatrix}, \\ g_2(t) &= \begin{pmatrix} \frac{1}{c^2}b \\ -\frac{1}{c}(e_2 + \frac{1}{c}ba_2) \\ e_1 + \frac{1}{c^2}be_2 - c(a_1 - \frac{a_1^3}{3} + a_2) + \frac{b}{c^3}(a_1 - a + ba_2) \end{pmatrix}. \end{aligned} \quad (4.3)$$

有了 A 、 g_1 和 g_2 , 即可得到诊断函数 $p_i(t) = \alpha_i^T g_i(t)$ ($i = 1, 2$) 的表达式, 其中 α_i 为 Φ_i 最大特征根对应的特征向量, 而 Φ_i 由 (2.23) 所定义.

下面考虑 $b(t)$ 的扰动分析. 由 (2.26) 可知, 对于以上模型, 状态变量 $V(t)$ 和 $R(t)$ 对应的 $g(t)$ 表达式分别为

$$g_1(t) = \begin{pmatrix} 0 \\ 0 \\ a_1 - \frac{a_1^3}{3} + a_2 \end{pmatrix}, \quad g_2(t) = \begin{pmatrix} \frac{1}{c} \\ -\frac{a_2}{c} \\ \frac{a_1 - a + ba_2}{c^2} \end{pmatrix}. \quad (4.4)$$

同理, 即可得到诊断函数 $p_i(t) = \alpha_i^T g_i(t)$, $i = 1, 2$, 其中 α_i 同上.

对于 $w(t)$ 的扰动分析, 由 (2.28) 可知, $g(t)$ 函数表达式为

$$g(t) = \begin{pmatrix} \frac{e_2}{c} \\ -\frac{a_2 e_2}{c} \\ (a_1 - \frac{a_1^3}{3} + a_2)e_1 + \frac{e_2^2}{c^2}(a_1 - a + ba_2) \end{pmatrix}. \quad (4.5)$$

由此即可得到诊断函数 $p_i(t) = \alpha_i^T g_i(t)$, $i = 1, 2$, 其中 α_i 为 Φ_i 最大特征根对应的特征向量, 而 Φ_i 由 (2.23) 所定义.

图 1-3 分别给出了 $a(b)$ 、 $b(t)$ 和 $w(t)$ 及其对应的诊断函数的图形. 作为测试, 我们在第 15 个观测值处加上了 1.5 倍的扰动. 由图可知, $p(t)$ 有三个重要特征: (1) 边界区域对参数估计的影响大; (2) 拐

图 1 对观测值 V 第 15 个点添加 1.5 倍的扰动后 V 、 R 分别对应的 $p(t)$ 曲线 图 2 对 V 的导数第 15 个点添加 1.5 倍的扰动后 V 、 R 分别对应的 $p(t)$ 曲线 图 3 权重函数第 15 点加 1.5 倍的扰动后对应的 $p(t)$ 曲线

点处的区域对参数估计影响大; (3) 不同区域对扰动反应不同. 实际上, 就我们对其他常微分方程的局部影响分析发现, 边界区域通常对参数估计都具有强影响, 这与已有的研究结果是一致的. 周杰等^[19]和 Zhou^[20]基于两步估计, 利用数据删除模型研究了常微分方程强影响点的检测, 同样也发现了类似

的现象. 不过相对于已有的研究, 目前的方法更简单, 易于实现. 对这种所谓的边界效应, 我们认为主要是由非参数回归的边界效应引起的, 在参数的两步估计中应该通过选择适当的权重函数 $w(t)$, 降低边界区域对于参数估计的影响. Gugushvili 和 Klaassen^[13] 理论上证明了, 若在两步估计中使用权重函数降低边界区域的影响, 则两步估计可达到 $\sqrt{n}$ 的收敛速度, 也即边界区域对于两步估计有着负面作用. 而这里的诊断函数 $p(t)$ 则表明, 边界区域不仅仅具有负面作用, 而且相对内部区域, 对参数估计有着更高的影响度. 诊断函数给出了不同区域对参数估计影响的直观展示, 为两步估计中权重函数 $w(t)$ 的选择提供了可行的基础.

5 实例分析

本节把上述方法应用到微生物研究中的一个实际数据集上 (参见文献 [21]). 此次实验的目的是利用高维的 ODE (ordinary differential equation) 模型建立老鼠肠道细菌之间的网络结构. 研究者把正常条件饲养的成年老鼠盲肠内容物移植到五个实验无菌老鼠体内, 对此五个实验老鼠的排泄物内的 17 类细菌种群 (operational taxonomic units, OTUs) 的相对丰度, 连续观测 20 天, 并用下述 17 维的 Lotka-Volterra 种间竞争方程组描述此时间序列数据动态特征:

$$\frac{dX_i(t)}{dt} = \alpha_i X_i(t) \left(1 - \frac{X_i(t)}{K} \right) + X_i(t) \left(\sum_{j=1, j \neq i}^{17} \beta_{ij} X_j(t) \right), \quad i = 1, 2, \dots, 17, \quad (5.1)$$

其中 $X_i(t)$ 表示第 i 个 OTU 的实际相对丰度. 观测值是在 $i = 1, \dots, 20$ 时, $X_i(t)$ 带噪声的测量值. 在这个 17 维的 ODE 模型中共有 289 个未知参数. 下面利用上文提出的统计诊断方法得到诊断函数 $p(t)$. 由于五只老鼠的实验结果相似, 这里仅给出标记为 #337D 的老鼠状态估计的诊断函数 $p(t)$ 曲线, 其中红色曲线代表状态估计, 蓝色曲线代表诊断函数.

由图 4 可知, 大部分诊断函数在边界区域取值较大, 这与模拟数据的研究是一致的, 也即两步估

图 4 老鼠肠道细菌 ODE 方程组 17 个 OTU 对应的诊断函数. 图中, 横坐标表示时间, 纵坐标表示细菌个数

计具有边界效应. 同时, OTU4、OTU10 和 OTU13 所对应的诊断函数在内部区域具有较高的取值, 说明这些区域对参数估计有较大的影响, 应给予更多的关注和分析.

6 结论

本文对因变量和自变量均为函数型数据的回归模型, 提出了一种简单易行的统计诊断方法, 并把此方法应用到常微分方程的统计诊断问题中, 数值分析显示所给出的诊断函数可以有效发现强影响区域. 特别地, 我们发现了常微分方程两步估计中的所谓的边界效应, 与此前已有的研究相吻合, 说明新方法研究与已有研究是一致的, 也为改进常微分方程的参数估计提供了可能性, 例如, 可以利用统计诊断函数设计更好的权重函数, 从而改进常微分方程的两步估计等. 必须指出的是, 鉴于函数型回归模型的复杂性, 如何用一种更科学全面的方法利用本文提出的诊断函数, 以提高参数估计效率, 还有待进一步的研究.

参考文献

- 1 Ramsay J O. When the data are functions. *Psychometrika*, 1982, 47: 379–396
- 2 Jolliffe I. *Principal Component Analysis*, 2nd ed. New York: Springer, 2002
- 3 Morris J S. Functional regression. *Annu Rev Stat Appl*, 2015, 2: 321–359
- 4 Arribas-Gil A, Müller H G. Pairwise dynamic time warping for event data. *Comput Statist Data Anal*, 2014, 69: 255–268
- 5 Muller H G. Functional modelling and classification of longitudinal data. *Scand J Statist*, 2005, 32: 223–240
- 6 Ramsay J O, Silverman B W. *Functional Data Analysis*, 2nd ed. Beijing: Science Press, 2005, 307–348
- 7 Boente G, Salibián-Barrera M. *S*-estimators for functional principal component analysis. *J Amer Statist Assoc*, 2015, 110: 1100–1111
- 8 Davies P L. Asymptotic behaviour of *S*-estimates of multivariate location parameters and dispersion matrices. *Ann Statist*, 1987, 15: 1269–1292
- 9 Bali J L, Boente G, Tyler D E, et al. Robust functional principle components: A projection-pursuit approach. *Ann Statist*, 2012, 39: 2852–2882
- 10 Crambes C, Delsol L, Laksaci A. Robust nonparametric estimation for functional data. *J Nonparametr Stat*, 2008, 20: 573–598
- 11 Huber P J. Robust estimation of a location parameter. *Ann Math Statist*, 1964, 35: 73–101
- 12 Huber P J. *Robust Statistics*. New York: John Wiley & Sons, 1981
- 13 Gugushvili S, Klaassen C A J. $\sqrt{n}$ -consistent parameter estimation for systems of ordinary differential equations: Bypassing numerical integration via smoothing. *Bernoulli*, 2012, 18: 1061–1098
- 14 Liang H, Wu H. Parameter estimation for differential equation models using a framework of measurement error in regression models. *J Amer Statist Assoc*, 2008, 103: 1570–1583
- 15 Wu W, Srivastava A. Analysis of spike train data: Alignment and comparisons using the extended Fisher-Rao metric. *Electron J Statist*, 2014, 8: 1776–1785
- 16 FitzHugh R. Impulses and physiological states in theoretical models of nerve membrane. *Biophys J*, 1961, 1: 445–466
- 17 Nagumo J, Arimoto S, Yoshizawa S. An active pulse transmission line simulating nerve axon. *Proc IRE*, 1962, 50: 2061–2070
- 18 Fan J, Gijbels I. *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall/CRC, 1996
- 19 周杰, 刘三阳, 周芳. HIV 模型的统计诊断. *科学通报*, 2012, 57: 666–672
- 20 Zhou J. Detection of influential measurement for ordinary differential equation with application to HIV dynamics. *Statist Probab Lett*, 2015, 107: 324–332
- 21 Marino S, Baxter N T, Huffnagle G B, et al. Mathematical modeling of primary succession of murine intestinal microbiota. *Proc Natl Acad Sci USA*, 2014, 111: 439–444

Local perturbation analysis of functional regression with applications

Jie Zhou, Zelin Wei & Liming Hu

Abstract For a special functional regression model, i.e., both the dependent and independent variables are functions, we propose the concept of continuous-time perturbation analysis. For three perturbation schemes which correspond to the independent variable, dependent variable and weight function respectively, we obtain the closed form of the diagnosis function based on the functional optimization theory. As an application, we applied this approach to the system of ordinary differential equations (ODEs) and discovered the so-called boundary effect of two-step estimation of ODEs which leaves the possibility for future improvement of two-step estimation.

Keywords functional regression model, nuisance parameters, local perturbation analysis, diagnosis function, boundary effect

MSC(2010) 62G10, 62H11

doi: 10.1360/N012017-00047