УДК 62-50

ПАВЛОВ А.А.,
КАЛАШНИК В.В.,
ХАЛИМОН А.Ю.

ПОСТРОЕНИЕ ОДНОМЕРНОЙ НЕЛИНЕЙНОЙ РЕГРЕССИИ НА ОСНОВЕ СПЛАЙН-ТЕХНОЛОГИИ И НОРМИРОВАННЫХ ОРТОГОНАЛЬНЫХ ПОЛИНОМОВ ФОРСАЙТА

В статье приводится метод восстановления одномерной нелинейной регрессии на заданном отрезке изменения значений аргумента, с произвольно распределенной аддитивной помехой, верхняя оценка дисперсии которой считается известной. В качестве сплайн-функций выбираются степенные полиномы, коэффициенты которых находятся как с помощью нормированных ортогональных полиномов Форсайта на основе теоретических и экспериментальных исследований, приведенных в [2], так и с помощью специально сконструированных задач линейного программирования. В завершении приводится обоснованный конструктивный критерий, гарантирующий точное решение поставленной задачи.

The article describes method of univariate non-linear regression recovery on given range of argument value variation, with arbitrary distributed additional noise, upper estimate of which is considered to be known. Power polynomials are chosen as spline functions, which coefficients can be found using Forsyth normalized orthogonal polynomial based on theoretical and practical research, given in [2], or using special linear programming models. Also, substantiated constructive criterion, that ensures accurate solution of formulated problem, is given.

Введение

Проблема построения нелинейной регрессии по зашумленным (с произвольным распределением) данным является одной из наиболее практически востребованных задач, которая даже в одномерном случае представляет в настоящее время достаточный теоретический и практический интерес [1,2].

Постановка задачи

Регрессионная модель имеет вид

$$Y(x) = \varphi(x) + \Delta,\ M\Delta = 0, \qquad (1)$$

где $\varphi(x)$ неизвестная (предполагается непрерывная и непрерывно дифференцируемая функция), $\Delta-$ произвольно распределённая случайная величина, верхняя оценка дисперсии которой σ^2 – известна, x – скалярный действительный аргумент функции $\varphi(x)$. Экспериментальные данные представлены в следующем виде: на отрезке [a,b] для значений аргумента x – $a = x_1, x_2, ..., x_n = b$ заданы числа y_i, $i = \overline{1,n}$. $y_i = \varphi(x_i) + \delta_i$, (2) δ_i – реализации случайной величины Δ. Предполагается, что нахождение практически точных значений $\varphi(x)$ $i = \overline{1,n}$ с достаточной для инженерной практики точностью задает функцию $\varphi(x)$ на отрезке [a,b].

Метод решения задачи

Используя сплайн-технологию отрезок [a,b] разбиваем на отрезки B_j, $j = \overline{1,m}$; $B_j = [a_j, b_j]$, $a_1 = a, b_j = a_{j+1}, j = \overline{1,m-1}, b_m = b,$.
$b_j \in \{x_1, ..., x_n\} = I_n$. Каждый отрезок B_j содержит n_j чисел из множества I_n. На B_j отрезке функция $\varphi(x)$ аппроксимируется полиномом степени m_j, $m_j < n_j$.

$$\varphi(x) \approx a_{0j} + a_{1j}x + ... + a_{m_j j}x^{m_j} \quad (3)$$

Для нахождения оценок $a_{ij}, i = \overline{1 \vee 2, m_j}$ используем нормированные ортогональные полиномы Форсайта [2]. Как показано в [2] (гл. 6) для σ^2 не превышающей 10^4 и $a \geq 40$ для гарантировано точной оценки коэффициентов a_{ij}, $i = \overline{2, m_j}$ достаточно взять $n_j \geq 10$ (в предположении что (3) точно описывает $\varphi(x)$ на отрезке B_j). Более того, [2] на примерах показало, что для $n_j = 10$ a_{1j} находится с точностью до второго знака после запятой. Критерием правильности построения отрезков B_j, $j = \overline{1,m}$ является существование натуральных чисел $m_j^1 < m_j$, $j = \overline{1,m}$ для которых оценки $a_{lj}, l = \overline{m_j^1 + 1, m_j}$, $j = \overline{1,m}$ практически являются оценками нулей.

Если это условие не выполняется, необходимо увеличить количество экспериментов на отрезке [a,b].

Таким образом аппроксимирующие полиномы на отрезках B_j, $j=\overline{1,m}$ построены с точностью до значений a_{0j}, либо a_{0j}, a_{1j}, $j=\overline{1,m_j}$.

Потребуем, чтобы в точках $b_j = x_{(j)}$, $j=\overline{1,m-1}$ значения апроксимирующих полиномов и их первых производных совпадали с заданной точностью. Для этого решим следующие задачи линейного программирования.

Модель 1

$$\min \sum_{i=1}^{n} z_i \tag{4}$$

$$-\varepsilon_1 \le \sum_{l=0}^{m_i^1} a_{li} x_{(i)}^l - \sum_{l=0}^{m_{i+1}^1} a_{l(i+1)} x_{(i)}^l \le \varepsilon_1 \tag{5}$$
$$i=\overline{1,m-1}$$

$$-\varepsilon_2 \le \sum_{l=1}^{m_i^1} l a_{li} x_{(i)}^{l-1} - \sum_{l=1}^{m_{i+1}^1} l a_{l(i+1)} x_{(i)}^{l-1} \le \varepsilon_2 \tag{6}$$
$$i=\overline{1,m-1}$$

$$-\phi_0^{-1}\left(\frac{1-\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \le \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=0}^{m_{j_i}^1} a_{lj_i} x_i^l\right) \le \phi_0^{-1}\left(\frac{1-\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \tag{7}$$

$$-z_i \le y_i - \sum_{l=0}^{m_{j_i}^1} a_{lj_i} x_i^l \le z_i, i=\overline{1,n} \tag{8}$$
$$z_i \ge 0, i=\overline{1,n}$$

где j_i – номер B_j отрезка, j_i равняется j если x_i принадлежит отрезку B_j.

Переменными задачи (4)-(8) являются $a_{0i}, a_{1i}, i=\overline{1,m}, z_i=\overline{1,n}$.

ϕ_0 – функция Лапласа, $\alpha \le 0{,}05$; $\varepsilon_1 \ge 0, \varepsilon_2 \ge 0$ – заданные константы.

Ограничения (7) с вероятностью $1-\alpha$ реализуют проверку статистической гипотезы о том, что полученные по восстановленной регрессии $\hat{\varphi}(x)$ числа $\hat{\delta}_i$ являются реализациями случайной величины Δ, где среднее арифметическое на основании теоремы Муавра-Лапласа имеет нормальное распределение.

Модель 2

Отличается от модели 1 тем, что в ней переменными являются $a_{0i}, i=\overline{1,m}; z_j, j=\overline{1,n}$. Модель 2 используется когда оценка $a_{1i}, i=\overline{1,m}$ с помощью нормированных ортогональных полиномов является достаточно точной.

Модель 3

Отличается от модели 1 тем, что переменными являются $a_{lj}, l=\overline{0,m_j^1}$; $j=\overline{1,m}$; $z_i, i=\overline{1,n}$.

Модель 4

$$\min \sum_{i=1}^{n} z_i \tag{9}$$

$$-\varepsilon_1 \le a_{0i} + a_{1i}x_{(i)} + \sum_{l=2}^{m_i^1}(a_{li}+\hat{a}_{li})x_{(i)}^l - a_{0(i+1)} - a_{1(i+1)}x_{(i)} - \sum_{l=2}^{m_i^1+1}(a_{l(i+1)}+\hat{a}_{l(i+1)})x_{(i)}^l \le \varepsilon_1 \tag{10}$$

$$-\varepsilon_2 \le a_{1i}x_{(i)} + \sum_{l=2}^{m_i^1} l(a_{li}+\hat{a}_{li})x_{(i)}^{l-1} - a_{1(i+1)}x_{(i)} - \sum_{l=2}^{m_i^1} l(a_{l(i+1)}+\hat{a}_{l(i+1)})x_{(i)}^{l-1} \le \varepsilon_2 \tag{11}$$

$$-\Phi^{-1}\left(\frac{1-\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \le \frac{1}{n}\sum_{i=1}^{n} y_i - a_{0j_i} - a_{1j_i}x_i - \sum_{l=2}^{m_{j_i}^1}(a_{lj_i}+\hat{a}_{lj_i})x_i^l \le \Phi^{-1}\left(\frac{1-\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \tag{12}$$

$$-z_i \le y_i - a_{0j_i} - a_{1j_i}x_i - \sum_{l=2}^{m_{j_i}^1}(a_{lj_i}+\hat{a}_{lj_i})x_i^l \le z_i, z_i \ge 0, i=\overline{1,n}. \tag{13}$$

$$-\varepsilon_3 \mid a_{lj} \mid \le \hat{a}_{lj} \le \varepsilon_3 \mid a_{lj} \mid, l=2, m_j^1, j=\overline{1,m} \tag{14}$$

Переменными задачи (9)-(14) являются a_{0j}, $a_{1j}, \hat{a}_{lj}, l=2, m_j^1, j=\overline{1,m}$, $z_i, i=\overline{1,n}$; $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, $\varepsilon_3 > 0$ – заданные погрешности.

В результате решения задач линейного программирования (4)-(14) получаем оценки неизвестной линии регрессии $\varphi(x)$, $\hat{\varphi}^l(x), l=\overline{1,4}$. На сегменте $[a,b]$ необходимо обоснованно выбрать наилучшую аппроксимацию. Предлагается сле-

дующая процедура выбора наилучшей аппроксимации из претендентов $\hat{\varphi}^l(x), l=\overline{1,4}$.

По каждой функции $\hat{\varphi}^l(x)$ находим оценки $\hat{\delta}_i^l, i=\overline{1,n}$, реализаций δ_i случайной величины Δ.

Случайным образом, имитирующим независимые испытания над Δ моделируем искусственным образом заданное число экспериментов $y_i^{lj}, x_i,\ i=\overline{1,n},\ \ j=\overline{1,k},\ \ k$ – количество искусственно конструируемых экспериментов $y_i^{lj}=\hat{\varphi}^l(x_i)+\hat{\delta}_i^{lj}, i=\overline{1,n}, j=\overline{1,k}$. Числа $\hat{\delta}_i^{lj}$ каждый раз случайным образом выбираются в каждом j-том ($j=\overline{1,k}$) искусственном эксперименте из множества $\{\hat{\delta}_i^l\}$.

По l-той задаче линейного программирования находим $\hat{\varphi}^{lj}(x), j=\overline{1,k}$, аппроксимирующий $\varphi(x)$ на отрезке $[a,b]$ ($\hat{\varphi}^{lj}(x)$ - j-тая аппроксимация $\varphi(x)$ построенная по j-ому эксперименту с помощью аппроксимации $\hat{\varphi}^l(x)$.

Построим множество чисел $\left\{\left|\hat{\varphi}^l(x_i)-\hat{\varphi}^{lj}(x_i)\right|, i=\overline{1,n}, j=\overline{1,k}\right\}=\gamma_{lji}\left|\hat{\varphi}^l(x_i)\right|$ и пусть γ_l является максимальным из них $\max\gamma_l=\max\limits_{ji}\gamma_{lji}$. Тогда в качестве наилучшей апроксимации выбирается $\hat{\varphi}^l(x)$, на которой достигается $\min\limits_l\gamma_l$

Выводы: Предложенные алгоритмы оценки неизвестной линии регрессии на заданном отрезке $[a,b]$ используют возможности нормированных ортогональных полиномов [2], сплайн-технологию, универсальность моделей линейного программирования. Эффективность метода следует из того, что он включает в себя три контрольные процедуры:

степень каждого сплайн-полинома меньше числа значений аргумента $\varphi(x)$;

каждая задача линейного программирования включает в себя статистический критерий проверки;

статистически является обоснованным, что число $\min\limits_l\gamma_l=\gamma_t$ может быть небольшим только в случае, когда множество $\{\hat{\delta}_i^t\}$ действительно является множеством реализаций случайной величины Δ, а это возможно, когда $\hat{\varphi}^t(x)$ практически точно оценивает $\varphi(x)$ в точках $x_i,\ i=\overline{1,n}$.

Задача имеет такое же решение, если модель представляется в виде

$$Y(t)=\varphi(t)+\Delta(t), \qquad (15)$$

Где t– действительный аргумент, который интерпретируется как время , $\Delta(t)$-стационарный в узком смысле слова случайный процесс, у которого для $\forall n,t_1,...,t_n$, n-мерная функция распределения $F(t_1,...,t_n)=\prod\limits_{j=1}^{n}F(t_j),\ F(t_j)$ – функция распределения случайной величины $\Delta(t_j)$.

Список литературы

1. Норман Р. Дрейпер Прикладной регрессионный анализ // Норман Р. Дрейпер, Гарри Смит. – Москва, Санкт-Петербург, Киев:, 2007. – 911с.
2. Згуровский М.З., Павлов А.А. Принятие решений в сетевых системах с ограниченными ресурсами: Монография.– К.: Наукова думка, – 2010.– 573 с.