

Г. М. Кошкин, В. А. Симахин и др. ; Тос. гос. ун-т. Томск, 1974.

3. Шуленин В. П. Введение в робастную статистику / Тос. гос. ун-т. Томск, 1993.

4. Воинов В. Г., Никулин М. С. Несмещенные оценки и их применения. М. : Наука, 1989.

5. Шурыгин А. М. Прикладная статистика. Робастность. Оценивание. Прогноз. М. : Финансы и статистика, 2000.

6. Beran R. An efficient and robust adaptive estimator of location // Ann. Stat. 1978. Vol. 6, № 2. P. 292–313.

7. Симахин В. А. Непараметрическая статистика. Ч. II. Теория оценок / Курган. гос. ун-т. Курган, 2004.

8. Симахин В. А. Взвешенный метод максимального правдоподобия // Высокие технологии XXI века : материалы IX Междунар. науч.-техн. конф. : в 2 т. Т. 2. Воронеж, 2008. С. 661–672.

9. Васильев В. А., Добровидов А. В., Кошкин Г. М. Непараметрическое оценивание функционалов от распределений стационарных последовательностей. М. : Наука, 2004.

V. A. Simakhin

ROBUST NONPARAMETRIC ESTIMATION OF LINEAR FUNCTIONALS

Robust nonparametric algorithms for estimation of linear functionals on the basis of weighted maximum likelihood method is considered in the article.

Keywords: robust, nonparametric, linear functional.

© Симахин В. А., 2010

УДК 62-506.1

Н. А. Сергеева, Е. С. Терентьева

О НЕПАРАМЕТРИЧЕСКИХ ОЦЕНКАХ ФУНКЦИИ РЕГРЕССИИ И ЕЕ ПРОИЗВОДНЫХ ПРИ НАЛИЧИИ ПРОПУСКОВ ДАННЫХ

Рассмотрены непараметрические методы оценивания регрессии и ее производных по выборкам случайных величин с некоторыми особенностями при их измерении. Представлен бутстреп-метод, применяемый для решения задачи заполнения пропусков в неполных данных или устранения пустот в пространстве наблюдений.

Ключевые слова: непараметрическая оценка регрессии, H -аппроксимация, бутстреп-метод, непараметрическая оценка производной функции регрессии, сходимость оценок.

Проблема моделирования дискретно-непрерывных процессов является одной из центральных в кибернетике. Определяющее значение при постановке задачи идентификации имеет математическая постановка, соответствующая различным априорным предположениям. Априорные сведения о процессе, по существу, определяют подход к задаче идентификации.

Ниже мы остановимся на задаче идентификации и связанной с ней задаче оценивания соответствующих вероятностных характеристик в условиях непараметрической неопределенности. В отличие от ставшего традиционным параметрического подхода к решению задачи идентификации в дальнейшем нам понадобятся некоторые качественные свойства поведения исследуемого процесса. Одним из главных этапов на пути решения этой задачи является оценивание регрессионных характеристик входных-выходных переменных процесса.

Непараметрический уровень априорной информации не предполагает наличия этапа выбора параметрической структуры модели, но требует некоторых сведений качественного характера о процессе, например от однозначности или неоднозначности его ха-

рактеристик, линейности для динамических процессов или характере нелинейности. При идентификации линейных динамических объектов мы сталкиваемся с необходимостью оценивания производной функции регрессии. Это связано с оценкой весовой функции линейной системы по измерениям функции переходной характеристики последней. Непараметрическая модель в этом случае представляет собой оценку интеграла Дюамеля.

Существенная особенность данного исследования состоит в предположении, что исходные выборки содержат пропуски данных при контроле входных-выходных переменных объекта. Это приводит к необходимости построения модифицированных непараметрических оценок функции регрессии и ее производных.

Пусть имеется неравномерная выборка статистически независимых наблюдений (u_i, x_i) , $i = \overline{1, s}$, входных и выходных переменных системы объемом s . Здесь u_i – значение вектора наблюдений входных воздействий размерности m в i -й точке выборки, а x_i – значение выходного воздействия в этой точке. Требу-

ется построить непараметрическую модель объекта и восстановить производную первого порядка стохастической зависимости $x(u)$ по имеющейся выборке наблюдений.

На начальном этапе восстановления $x(u) \forall u \in \Omega(u)$ принимается статистика [1]:

$$\tilde{x}(u) = \sum_{j=1}^s x_j \prod_{k=1}^m \Phi\left(\frac{u^k - u_j^k}{C_s^k}\right) / \sum_{j=1}^s \prod_{k=1}^m \Phi\left(\frac{u^k - u_j^k}{C_s^k}\right), \quad (1)$$

где $\Phi(z)$ – колоколообразная функция, удовлетворяющая следующим условиям [1]:

$$\begin{aligned} \Phi(z) < \infty, \forall z \in \Omega(z), \quad \int_{\Omega(z)} \Phi(z) dz = 1, \\ \int_{\Omega(z)} \Phi^2(z) dz < \infty, \quad \lim_{s \rightarrow \infty} \frac{1}{C_s} \cdot \Phi\left(\frac{u^j - t^j}{C_s}\right) = \delta(u^j - t^j), \end{aligned} \quad (2)$$

здесь $\delta(u^j - t^j)$ – дельта-функция Дирака, z – аргументы колоколообразной функции.

Параметр размытости C_s должен удовлетворять условиям [1]:

$$C_s > 0, \quad \lim_{s \rightarrow \infty} C_s = 0, \quad \lim_{s \rightarrow \infty} s \cdot C_s^m = \infty. \quad (3)$$

На следующем этапе строится H -аппроксимация [2]:

$$x_k(u) = \frac{1}{kc_k^m} \sum_{i=1}^k x_i^* \prod_{j=1}^m H\left(\frac{u^j - u_i^{*j}}{c_k}\right), \quad (4)$$

где $\{u_i^*, x_i^*\}$ – рабочая равномерная выборка объемом $k < s$; $c_k^m \rightarrow 0$; $kc_k^m \rightarrow \infty$; $\forall j = 1, \dots, m$. Тогда

$\frac{1}{c_k} H\left(\frac{u^j - t^j}{c_k}\right) \rightarrow \delta(u^j - t^j)$ с ростом k , где $H(z)$ – колоколообразная функция, отличающаяся от $\Phi(z)$ на множитель, равный константе

$$\frac{1}{kc_k^m} \sum_{i=1}^k \prod_{j=1}^m \Phi\left(\frac{u^j - u_i^{*j}}{c_k}\right) = \text{const} \quad \forall u \in \Omega(u).$$

Можно показать, что

$$\frac{1}{kc_k^m} \sum_{i=1}^k \prod_{j=1}^m H\left(\frac{u^j - u_i^{*j}}{c_k}\right) = 1 \quad \forall u \in \Omega(u). \quad (5)$$

К классу функций $H(z)$ и $\Phi(z)$ относятся, например, гауссова кривая и функция Соболева.

В качестве оценки производной регрессии предлагается взять аппроксимацию, построенную на рабочей равномерной выборке, в виде [3]:

$$D^\alpha x_k(u) = \frac{1}{kc_k^m} \sum_{i=1}^k x_i^* \prod_{j=1}^m H^{(\alpha)}\left(\frac{u^j - u_i^{*j}}{c_k}\right), \quad (6)$$

где $\alpha = (\alpha_1, \dots, \alpha_m)$, $\alpha_j \in \{0, 1\}$, $j = \overline{1, m}$, если $\alpha_j = 1$, то по переменной u^j берется производная первого порядка, если $\alpha_j = 0$, то производная нулевого порядка; $H^{(\alpha)}(z)$ – функция, удовлетворяющая интегральным условиям, сформулированным ниже.

Если u – скалярная величина, т. е. $m = 1$, то оценка первой производной регрессии (6) при $\alpha = 1$ является асимптотически несмещенной и состоятельной при выполнении следующих условий:

$$\begin{aligned} \int_{R^1} H^{(\alpha)}(z) dz = 0, \quad c_k \int_{R^1} z H^{(\alpha)}(z) dz = -1, \\ \int_{R^1} H^{(\alpha)}(z) dz < \infty. \end{aligned} \quad (7)$$

Техника доказательства основывается на результатах [3].

Заметим, что обозначенные выше условия сужают класс функций, которые могли бы применяться в случае восстановления кривой регрессии. В частности,

гауссова кривая $H(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$ (рис. 1, а) имеет производную $H'(z)$ (рис. 1, б), которая сложна с вычислительной точки зрения, однако ее можно заменить кусочно-постоянным аналогом $H^{</>}(z)$ (рис. 1, в). Отметим, что время вычисления $H^{</>}(z)$ в 1,3 раза меньше, чем $H'(z)$.

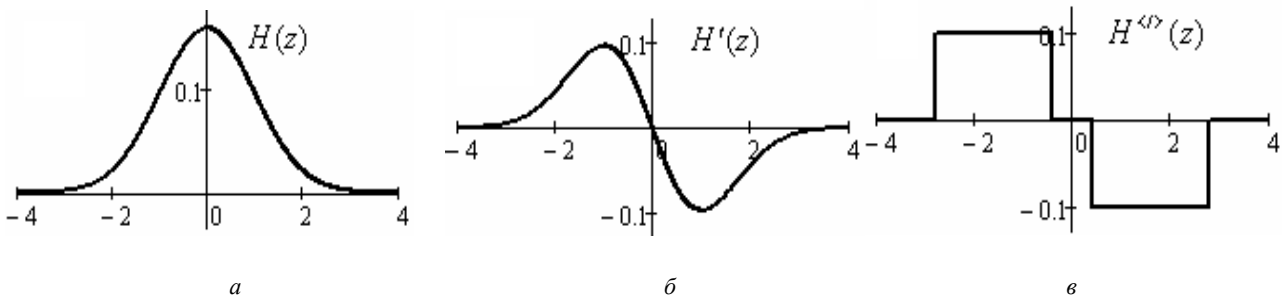


Рис. 1

Для выбора оптимального параметра размытости введем меру отклонения оценки (6) от истинной производной:

$$W(c_k) = \int_{G(u)} (Dx(u) - D^\alpha x_k(u))^2 du, \quad (8)$$

где $G(u)$ – область сравнения. Тогда критерий оптимальности будет иметь вид

$$W(c_k) \rightarrow \min_{\Omega(c_k)}, \quad (9)$$

где $\Omega(c_k) = (0, +\infty)$. Но при практическом применении данный критерий невозможно использовать, так как $D^\alpha x(u)$ неизвестна.

Преобразуем (9), используя формулу интегрирования по частям и тем самым понижая порядок производной. Тогда критерий оптимальности приобретет следующий вид:

$$\begin{aligned} I(c_k) = & \int_{G(u)} (D^\alpha x_k(u))^2 du - \\ & - 2 \int_{G(u^-)} \tilde{x}_k(u) D^\alpha x_k(u) \Big|_{G(u_j)} du + \\ & + 2 \int_{G(u)} \tilde{x}_k(u) D^{\alpha^+} x_k(u) du \rightarrow \min_{c_k}, \end{aligned} \quad (10)$$

где $u^- = (u_1, \dots, u_{j-1}, u_{j+1}, \dots, u_m)$, здесь j такое, что $\alpha_j \neq 0$, $\alpha^+ = (\alpha_1, \dots, \alpha_{j-1}, \alpha_j + 1, \alpha_{j+1}, \dots, \alpha_m)$; $\tilde{x}_k(u)$ – оценка любого типа, которая в смысле некоторого критерия достаточно хорошо аппроксимирует нужную нам величину; $G(u) \subset \Omega(u)$ – область хорошего качества оценки $\tilde{x}_k(u)$. Если это непараметрическая оценка, то вычисление оптимального параметра размытости становится рекуррентным. На первом шаге находится оптимальный параметр размытости c_k^0 для

оценки кривой регрессии. На следующем шаге определяется оптимальный параметр размытости c_k^1 для оценки производной кривой регрессии первого порядка по какой-либо переменной. Критерий оптимальности на этом шаге имеет вид (10) при $\tilde{x}(u) = \tilde{x}(u, c_k^0)$ и $\alpha = (0, \dots, 0, 1, 0, \dots, 0)$.

При построении модели реального процесса иногда используются выборки случайной величины, результаты измерений которой распределены неравномерно. Это приводит к тому, что в некоторых подобластях пространства наблюдений образуются пустоты. В таких условиях приходится отказываться от применения стандартной непараметрической оценки регрессии.

При наличии пустот в пространстве наблюдения $\Omega(u)$ непараметрическая оценка регрессии основывается на использовании не конкретного значения выходной переменной в j -й точке выборки, а ее оценки. Проводить это оценивание будем по представленному ниже алгоритму.

Пусть мы находимся в j -й точке выборки. Определим множество Λ_{sj} соседних точек выборки, в которых колоколообразная функция не равна нулю:

$$\Lambda_{sj} = \left\{ u_i : \prod_{k=1}^m \Phi \left(\frac{u_i^k - u_j^k}{C_s^i(c_s, \rho_s)} \right) > 0 \right\}, \quad (11)$$

$$i = \overline{1, s}, j = \overline{1, s},$$

где ρ_s – параметр оптимизации, а колоколообразная функция для каждой входной переменной u^k , $k = \overline{1, m}$, расширяется в направлении разрежений в выборке, т. е. ее ветви имеют разные константы Липшица. Приведем пример такой асимметричной функции:

$$\Phi \left(\frac{u^k - u_i^k}{C_s} \right) = \begin{cases} \frac{1}{2} \left(\cos \left(\frac{1}{l} \cdot \frac{u^k - u_i^k}{C_s} \right) \right) + \frac{1}{2}, & \text{if } -l \cdot \pi \leq \frac{u^k - u_i^k}{C_s} \leq 0 \\ \frac{1}{2} \left(\cos \left(\frac{u^k - u_i^k}{C_s} \right) \right) + \frac{1}{2}, & \text{if } 0 < \frac{u^k - u_i^k}{C_s} \leq \pi \\ 0, & \text{if } \left(\frac{u^k - u_i^k}{C_s} > \pi \right) \vee \left(\frac{u^k - u_i^k}{C_s} < -l \cdot \pi \right) \end{cases} \left. \vphantom{\begin{aligned} \frac{1}{2} \left(\cos \left(\frac{1}{l} \cdot \frac{u^k - u_i^k}{C_s} \right) \right) + \frac{1}{2}, \\ \frac{1}{2} \left(\cos \left(\frac{u^k - u_i^k}{C_s} \right) \right) + \frac{1}{2}, \\ 0, \end{aligned}} \right\} \text{if } f_s(u^k + h, c_s) > f_s(u^k - h, c_s),$$

$$\begin{cases} \frac{1}{2} \left(\cos \left(\frac{1}{l} \cdot \frac{u^k - u_i^k}{C_s} \right) \right) + \frac{1}{2}, & \text{if } 0 \leq \frac{u^k - u_i^k}{C_s} \leq l \cdot \pi \\ \frac{1}{2} \left(\cos \left(\frac{u^k - u_i^k}{C_s} \right) \right) + \frac{1}{2}, & \text{if } -\pi < \frac{u^k - u_i^k}{C_s} \leq 0 \\ 0, & \text{if } \left(\frac{u^k - u_i^k}{C_s} > l \cdot \pi \right) \vee \left(\frac{u^k - u_i^k}{C_s} < \pi \right) \end{cases} \left. \vphantom{\begin{aligned} \frac{1}{2} \left(\cos \left(\frac{1}{l} \cdot \frac{u^k - u_i^k}{C_s} \right) \right) + \frac{1}{2}, \\ \frac{1}{2} \left(\cos \left(\frac{u^k - u_i^k}{C_s} \right) \right) + \frac{1}{2}, \\ 0, \end{aligned}} \right\} \text{if } f_s(u^k + h, c_s) < f_s(u^k - h, c_s),$$

где $h > 0$ – радиус окрестности текущей точки u ; $l \geq 1$ – коэффициент расширения колоколообразной функции, при $l = 1$ колоколообразная функция принимает симметричный вид. Функция множества в виде непараметрической оценки плотности [4] с малым параметром размытости

$$f_s(u^k, c_s) = \frac{1}{sc_s} \sum_{i=1}^S \Phi\left(\frac{u^k - u_i^k}{c_s}\right), \quad k = \overline{1, m}, \quad (12)$$

применяется для определения сгущений и разрежений точек в выборке по каждой входной переменной u^k , $k = \overline{1, m}$. Форма и вид колоколообразной функции зависят от плотности точек выборки в h -окрестности текущей точки.

Рассмотрим выборку наблюдений входной переменной $u_i, i = \overline{1, 20}$ (рис. 2) с функцией множества (12) (рис. 3). Тогда график колоколообразной функции в точке $u_i = 0,26$ при разных значениях коэффициента расширения l будет иметь вид, представленный на рис. 4 колоколообразная функция расширилась по направлению уменьшения значения функции множества в окрестности текущей точки $u_i = 0,26$.

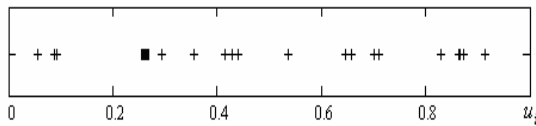


Рис. 2

+++ – точки, в которых производились измерения входной переменной объекта; ■■■ – $u_i = 0,26$

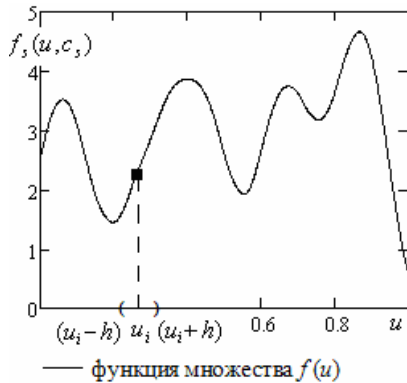


Рис. 3

На подобласти $\Omega(x, u)$, определяемой Λ_{sj} , сформированной в соответствии с правилом (11), строится поверхность $\phi_j(u) = \beta_j \Psi(u)$, $j = \overline{1, s}$, где $\Psi(u)$ – вектор базисных функций; β_j^T – вектор параметров, определяемых по методу наименьших квадратов.

Для примера рассмотрим выборку наблюдений со скалярным входным и выходным воздействиями $(u_i, x_i), i = \overline{1, 15}$ (рис. 5), где функции $\phi_j(u), j = \overline{1, 15}$, имеют линейный вид. На рис. 5 отмечена текущая де-

сящая точка выборки и точки вокруг нее, колоколообразная функция $\Phi((u - u_i)/C_s)$ в которых не равна 0. В данном случае их 4. Таким образом, вектор коэффициентов β_j^T при наборе базисных функций $\Psi(u) = (1, u)^T$ определяется по методу наименьших квадратов на основании пяти выделенных точек.

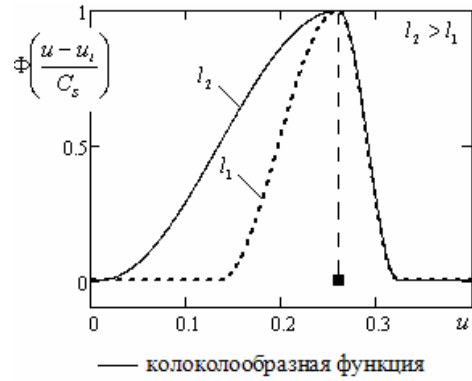


Рис. 4

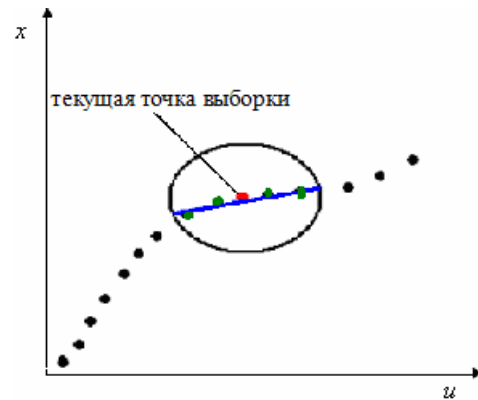


Рис. 5

Произведем модификацию непараметрической оценки регрессии следующим образом [5]:

$$\hat{x}(u, c_s, \rho_s) = \frac{\sum_{j=1}^S \left(\phi_j(u) \prod_{k=1}^m \Phi\left(\frac{u^k - u_j^k}{C_s^j(c_s, \rho_s)}\right) \right)}{\sum_{j=1}^S \prod_{k=1}^m \Phi\left(\frac{u^k - u_j^k}{C_s^j(c_s, \rho_s)}\right)}. \quad (13)$$

Параметр размытости в формуле (13) оценивается для каждой j -й точки из выборки наблюдений [6]:

$$C_s^j(c_s, \rho_s) = \rho_s \cdot f_s^{-1}(u_j, c_s), \quad j = \overline{1, s}. \quad (14)$$

Выбор оптимального параметра размытости c_s и коэффициента пропорциональности ρ_s осуществляется путем минимизации среднеквадратичного критерия:

$$W(c_s, \rho_s) = \frac{1}{s} \sum_{i=1}^s (x_i - \hat{x}(u_i, c_s, \rho_s))^2 \rightarrow \min_{c_s, \rho_s}. \quad (15)$$

Другим методом непараметрического оценивания регрессии между входными и выходными воздействиями объекта является предварительное заполнение пустот в пространстве наблюдений случайной величины. Для этого используется следующая бутстреп-процедура [7].

1. Точка u_i равномерной сетки, натянутой на область наблюдения $\Omega_s(u)$ входных переменных $u^k, k = \overline{1, m}$, будет считаться пропуском, если

$$\prod_{j=1}^m f_s(u_i^j, c_s) < a, a > 0. \quad (16)$$

Пусть таких точек будет K , т. е. $l = \overline{1, K}$. Параметр $a > 0$ настраивается исходя из оптимизационной процедуры по среднеквадратичному критерию.

2. По присутствующим наблюдениям $(u_i, x_i), i = \overline{1, s}$ входных и выходных переменных системы строится регрессионная модель (1), однако параметр размытости оценивается по формуле (14) для каждой точки выборки наблюдений, что позволяет избежать неадекватного поведения оценки (1) в областях разрежений (пустот) пространства наблюдений $\Omega_s(u)$.

3. По построенной регрессионной модели находят оценки $x_{si} = \tilde{x}(u_i), i = \overline{1, s}$.

4. Определяются ошибки $\varepsilon_i = x_{si} - x_i, i = \overline{1, s}$, для всех точек выборки.

5. Для каждого пропуска после подстановки значения сопутствующей входной переменной u_i в полученное регрессионное уравнение находится оценка $\tilde{x}(u_i), l = \overline{1, K}$.

6. Значения выходной переменной, которыми замещают пропуски, получается по формуле

$$x_{sl} = \tilde{x}(u_l) + \varepsilon_l, l = \overline{1, K}, \quad (17)$$

где ε_l выбирается случайно из ошибок, рассчитанных в п. 4. Это можно реализовать следующим образом: с помощью генератора случайных чисел выбирается целое число q на интервале $[1; s]$, $\varepsilon_l = \varepsilon_q$, и операция повторяется для каждого $\varepsilon_l, l = \overline{1, K}$.

7. Данные, полученные после заполнения пропусков, $(u_l, x_{sl}), l = \overline{1, K}$, объединяются с исходной выборкой наблюдений $(u_i, x_i), i = \overline{1, s}$. По итоговой выборке объема $s + K$ строится регрессионная модель (1).

Приведем численные результаты моделирования при использовании модифицированной непараметрической оценки регрессии (13) в сравнении со стандартной непараметрической оценкой (1).

Пусть размерность входной переменной равна двум и имеется неравномерная выборка наблюдений $(u_i, x_i), i = \overline{1, s}$, объемом $s = 200$, $u = (u^1, u^2)$, на области $\Omega_s(u) : \{u^1 \in [0; 0,5], u^2 \in [0; 0,5]\}$; $x = \varphi(u) + \xi$, где

$\varphi(u) = \varphi(u^1, u^2) = \sin(7 \cdot u^1 + u^2)$; ξ – аддитивная центрированная помеха, имеющая нормальный закон распределения с математическим ожиданием $M(\xi) = 0$ и ограниченной дисперсией.

Помеха накладывается следующим образом:

– измеряется интервал изменения сигнальной части $[a_1; a_2]$;

– задается уровень помех h на интервале $[0; 1]$;

– с помощью генератора случайных чисел формируется вектор ξ значений случайной величины, распределенной по нормальному закону с $M(\xi) = 0$,

$\sigma(\xi) = \frac{h(a_2 - a_1)}{6}$. Вектор ξ складывается с вектором значений сигнальной части.

Создадим пробел в выборке наблюдений, расположение точек $u_i, i = \overline{1, 200}$, для которой приведено на рис. 6. С помощью генератора случайных чисел формируем вектор $u^2, |u^2| = 200$, значений случайной величины, распределенной по равномерному закону на интервале $[0; 0,5]$. Затем генерируем векторы u_1^1 и $u_2^1, |u_1^1| = |u_2^1| = 100$, на интервалах $[0; 0,2]$ и $[0,3; 0,5]$ соответственно. Выборка наблюдений входных воздействий объемом 200 имеет вид $(u_1^1 \cup u_2^1; u^2)$.



Рис. 6

Пусть в каналах измерения выходного сигнала присутствует 5%-я помеха, т. е. $h = 0,05$, а $\phi_j(u) = \beta_{2j}u^2 + \beta_{1j}u^1 + \beta_{0j}, j = \overline{1, s}$, имеет линейный вид, где u^2, u^1 – входные переменные. Результат моделирования в виде среза ($u^2 = 0,25, u^1 = [0; 0,5]$) представлен ниже (рис. 7).

Среднеквадратичная оценка ошибки моделирования при использовании стандартной непараметрической оценки регрессии (1) равна 3,3 %. При использовании модифицированной оценки регрессии (13) с несимметричной колоколообразной функции оценка ошибки равна 1,5 %, что в 2 раза меньше, чем при ис-

пользовании стандартной оценки. Оценка ошибки моделирования здесь и в дальнейшем рассчитывалась следующим образом:

$$E = \frac{\sqrt{\sum_{i=1}^{100} (\varphi(u_i) - \hat{x}(u_i))^2}}{100} 100 \%,$$

где $u_i, i = \overline{1, 100}$ – вектор значений случайной величины, распределенной по равномерному закону на области $\Omega_s(u)$.

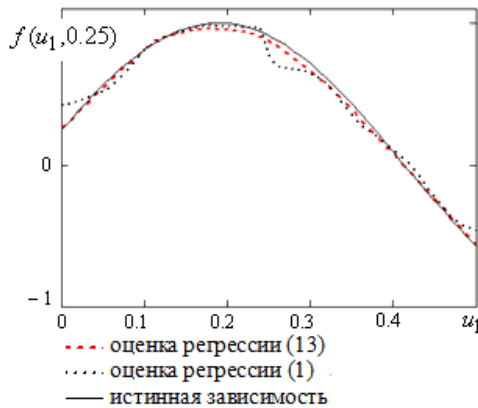


Рис. 7

Пусть размерность входной переменной равна четырем и имеется неравномерная выборка наблюдений $(u_i, x_i), i = \overline{1, s}$, объемом $s = 400$, $u = (u^1, u^2, u^3, u^4)$, на области $\Omega_s(u): \{u^1 \in [0; 0,5], u^2 \in [0; 0,5], u^3 \in [0; 0,5], u^4 \in [0; 0,5]\}$. И пусть в этой области имеются искусственно созданные пустоты по двум входным переменным: $u_i^3 \in [0; 0,2] \cup [0,25; 0,5]$, $u_i^4 \in [0; 0,3] \cup [0,4; 0,5]$, $i = \overline{1, 400}$; $x = \varphi(u) + \xi$, где ξ – аддитивная помеха, имеющая нормальный закон распределения с $M(\xi) = 0$ и $D(\xi) < \infty$, а истинная зависимость имеет вид

$$\varphi(u) = \varphi(u^1, u^2, u^3, u^4) = \sin(0,45 \cdot u^1) - \sin(0,5 \cdot u^2) + \sin(0,45 \cdot u^3) - \sin(0,5 \cdot u^4).$$

Пусть в каналах измерения выходного сигнала присутствует 5%-я помеха. Тогда результат моделирования при помощи модифицированной оценки регрессии (13) в виде среза ($u^1 = u^2 = 0,2, u^3 = 0,4, u^4 \in [0; 0,5]$) будет представлен в виде линейного аппроксимирующего полинома (рис. 8) и квадратичного аппроксимирующего полинома (рис. 9).

Оценка ошибки моделирования при использовании модифицированной оценки регрессии (13) с квадратичным аппроксимирующим полиномом $\hat{\varphi}_j(u), j = \overline{1, s}$, в 8,2 раза меньше, чем с линейным полиномом, и в 15,6 раз меньше, чем при использовании стандартной непараметрической оценки (1).

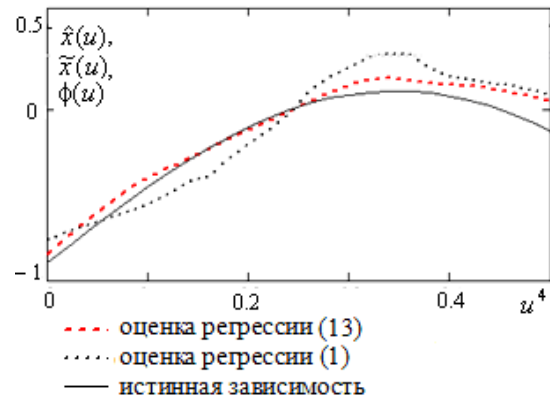


Рис. 8

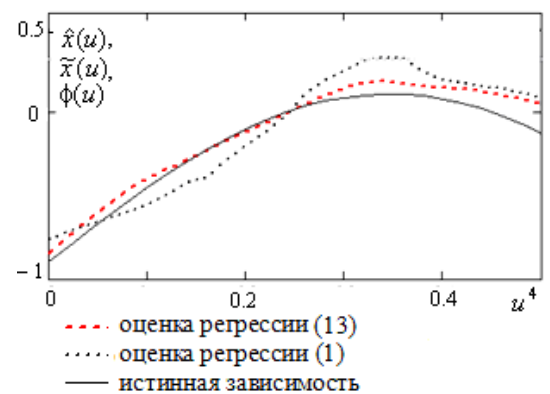


Рис. 9

Таким образом, результаты экспериментов, представленные на рис. 6–9, подтверждают, что использование модифицированной непараметрической оценки регрессии (13) с квадратичной аппроксимирующей функцией дает в несколько раз меньшую ошибку идентификации при наличии пустот в пространстве наблюдений, чем применение стандартной непараметрической оценки (1).

Выберем объект с одномерным входным и выходным воздействиями. Для этого примем $\varphi(u) = 5(u - 0,5)^2$. Пусть объем выборки равен $s = 30$: $(u_i, x_i), i = \overline{1, 30}$, а сама выборка наблюдений относительно равномерна и не имеет больших пробелов.

Результаты моделирования, приведенные на рис. 10, показывают, что среднеквадратичная оценка ошибки моделирования при использовании стандартной непараметрической оценки регрессии (1) в 1,5 раза больше, чем при применении оценки регрессии (13) с несимметричной колоколообразной функцией и линейным аппроксимирующим полиномом.

Таким образом, при разреженной выборке, без явных пробелов в пространстве наблюдений, ошибка моделирования с использованием модифицированной оценки регрессии (13) меньше, чем с применением стандартной оценки.

Проведем сравнение результатов численного моделирования, полученных с помощью модифициро-

ванной оценки регрессии (13) и бутстреп-метода со стандартной непараметрической оценкой (1).

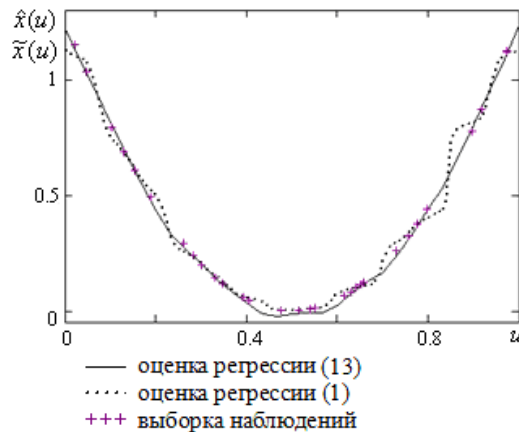


Рис. 10

Пусть имеется неравномерная выборка наблюдений одномерных входного и выходного воздействий объекта (u_i, x_i) , $i = \overline{1, s}$, объемом $s = 20$, $u \in [0; 1]$, $x = \varphi(u) + \xi$, где $\varphi(u) = \sin(5 \cdot u)$; ξ – 5 %-я помеха в каналах измерения выходного сигнала. Пусть $\phi_j(u) = \beta_{2j}u^2 + \beta_{1j}u + \beta_{0j}$, $j = \overline{1, s}$. Тогда результаты моделирования будут следующими (рис. 11, 12).

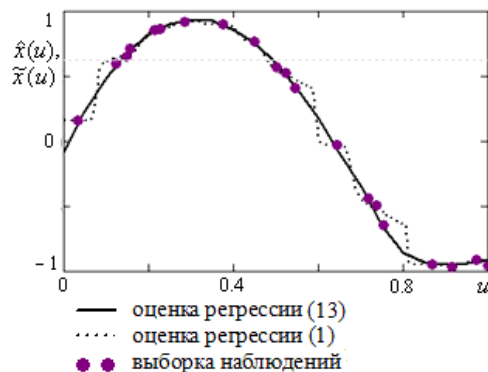


Рис. 11

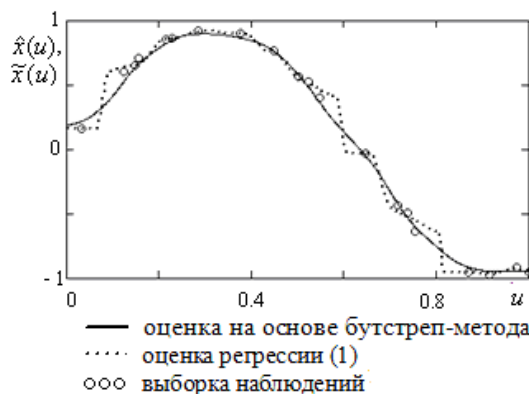


Рис. 12

Исходная выборка наблюдений входных воздействий может быть дополнена значениями наблюдений по бутстреп-методу (рис. 13).

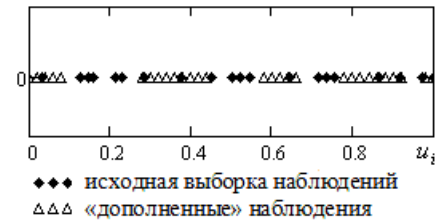


Рис. 13

Среднеквадратичная оценка ошибки моделирования при использовании стандартной непараметрической оценки регрессии (1) равна 3 %, при применении модифицированной оценки регрессии (13) – 1,45 %, а при построении непараметрической оценки (1) на основе выборки, сформированной по бутстреп-методу, – 1,5 % (рис. 14).

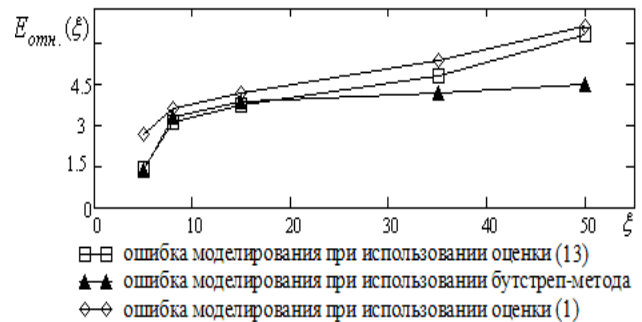


Рис. 14

Анализ графиков на рис. 14 позволяет сделать вывод, что при сравнительно малом уровне помехи в каналах измерения выходного сигнала (< 15 %) наименьшую ошибку моделирования дает модифицированная оценка регрессии (13). При увеличении уровня шума (> 15 %) ошибка моделирования на основе бутстреп-метода становится меньше, чем при использовании оценки (13). Следует отметить, что при высоком уровне помехи в каналах измерения (> 40 %) качество построения модели при помощи модифицированной оценки (13) становится практически эквивалентным качеству моделирования на основе стандартной непараметрической оценки (1).

Сравним результаты численного моделирования при использовании модифицированной оценки регрессии (13) в случае, когда функция локальной аппроксимации $\phi_j(u) = \beta_{2j}u^2 + \beta_{1j}u + \beta_{0j}$, $j = \overline{1, s}$, имеет квадратичный вид, и в случае, когда она линейна: $\phi_j(u) = \beta_{1j}u + \beta_{0j}$, $j = \overline{1, s}$.

Пусть имеется неравномерная выборка наблюдений одномерных входного и выходного воздействий объекта (u_i, x_i) , $i = \overline{1, s}$, объемом $s = 20$, $u \in (0; 1)$, $x = \varphi(u) + \xi$, где $\varphi(u) = \sin(5 \cdot u)$; ξ – 30%-я центрированная помеха в каналах измерения выходного сигнала.

Оценка ошибки моделирования при использовании непараметрической оценки регрессии (13) с квадратичным аппроксимирующим полиномом $\phi_j(u) = \beta_{2j}u^2 + \beta_{1j}u + \beta_{0j}$, $j = \overline{1, s}$, в 1,1 раза меньше, чем с линейным полиномом $\phi_j(u) = \beta_{1j}u + \beta_{0j}$, $j = \overline{1, s}$. При дальнейшем увеличении порядка полинома ошибка моделирования практически не изменяется, а при порядке полинома больше четырех ошибка моделирования увеличивается. Таким образом, предлагается использовать второй порядок аппроксимирующего полинома $\phi_j(u)$, $j = \overline{1, s}$.

Представим зависимости оценок ошибки моделирования при использовании стандартной непараметрической оценки (1) на основе исходной выборки наблюдений и выборки, дополненной по бутстреп-методу, а также модифицированной оценки регрессии с квадратичным аппроксимирующим полиномом от уровня шума, % (рис. 15).

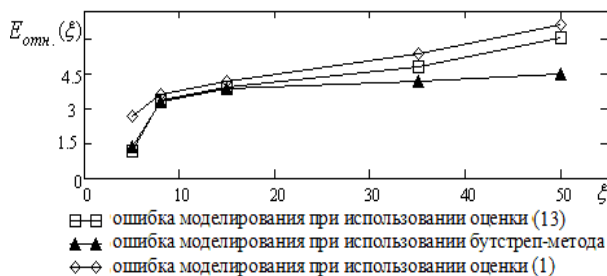


Рис. 15

Таким образом, разработанные методы построения непараметрических моделей позволяют довольно эффективно моделировать объекты в случае неравномерно распределенной выборки наблюдений входных и выходных воздействий. Применение несимметричной колоколообразной функции способствует большей согласованности модели и объекта по сравнению с симметричной функцией. Однако при увеличении размерности входной переменной на единицу время расчета модели при использовании модифицированной оценки регрессии (13) с несимметричной коло-

колообразной функцией увеличивается в среднем в 1,5 раза при фиксированном объеме выборки наблюдений.

В заключение дадим несколько рекомендаций по применению представленных оценок в различных условиях. В случае идентификации объекта на основе представительной выборки наблюдений можно использовать любую из предложенных оценок. При наличии пустот в пространстве наблюдений и малом уровне помехи в каналах измерения для моделирования больше подходит модифицированная оценка регрессии (13) с несимметричной колоколообразной функцией и квадратичным аппроксимирующим полиномом. Если в каналах измерения высокий уровень помех, то предлагается использовать непараметрическую оценку, построенную на основании дополненной по бутстреп-методу выборки наблюдений.

Библиографические ссылки

1. Надарая Э. А. Замечания о непараметрических оценках плотности вероятности и кривой регрессии // Теория вероятности и ее применение. 1970. Т. 15. Вып. 1. С. 139–142.
2. Медведев А. В. Непараметрические системы адаптации. Новосибирск : Наука. Сиб. отд-ние, 1983.
3. Медведева Н. А. Непараметрические оценки производной кривой регрессии и модели динамики // Информатика и процессы управления. Красноярск, 1995. С. 74–81.
4. Parzen E. On estimation of a probability density function and mode // Ann. Math. Stat. 1962. Vol. 33. P. 1065–1076.
5. Катковник В. Я. Непараметрическая идентификация и сглаживание данных. М. : Наука, 1985.
6. Непараметрическое моделирование стохастических систем / В. А. Гутшмидт, Я. И. Демченко, М. В. Кураченко и др. // Теория вероятностей, случайные процессы, математическая статистика и приложения. Минск, 2008. С. 66–73.
7. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных // Computer Modeling & New Technologies. 2002. Т. 6, № 1. С. 51–61.

N. A. Sergeeva, E. S. Terentyeva

NONPARAMETRIC ESTIMATIONS OF REGRESSION FUNCTION AND ITS DERIVATIVES IN THE PRESENCE OF DATA ADMISSIONS

In the article we consider nonparametric methods of estimation of a regression and its derivatives on samplings of random variables with some singularities at their measurement. A bootstrap-method applied to the decision of the passes filling task in incomplete data or elimination of emptiness in space of observations is presented.

Keywords: nonparametric estimation of a regression, H -approximation, a butstrep-method, nonparametric estimation of a derivative of a regression, convergence of estimations.

© Сергеева Н. А., Терентьева Е. С., 2010