УДК 519.234

В. А. Симахин

РОБАСТНЫЕ НЕПАРАМЕТРИЧЕСКИЕ ОЦЕНКИ ЛИНЕЙНЫХ ФУНКЦИОНАЛОВ

Рассматривается построение алгоритмов робастных непараметрических оценок линейных функционалов на основе взвешенного метода максимального правдоподобия.

Ключевые слова: робастный, непараметрический, оценка, линейный функционал.

Пусть $y_1, ..., y_M$ – выборка непараметрической оценки регрессии (НОР) с функцией распределения (ФР) $G(y)$ и $\theta = \int \varphi(\vec{t})dH(\vec{t}) < \infty$, где $\vec{t} = (t_1, ..., t_m)^T$; $H(\vec{t}) = G(t_1)\cdots G(t_m)$. Непараметрические оценки функционала θ при симметричных функциях $\varphi(\vec{t})$ получили название *U*-статистик [1; 2]. В классе робастных оценок θ применяется метод усечения выборки – усеченные *U*-статистики [3].

Обозначим через $f(x)$ и $F(x)$ плотность и ФР случайной величины $X = \varphi(Y_1, ..., Y_m)$, тогда $\theta = \int z dF(z)$. Выборку $Y_1, ..., Y_m$ преобразуем в выборку $x_1, ..., x_N$, где $x_j = \varphi(y_{i_1}, ..., y_{i_m})$; N – мощность множества $\{i_1 < i_2 < ... < i_m\}$. При таком преобразовании задача оценивания параметра θ сводится к задаче оценивания параметра сдвига распределения $F(x)$. В параметрической статистике такой прием широко используется для синтеза несмещенных оценок параметров как функций от достаточных статистик и в вычислительном отношении достаточно удобен, однако основная сложность здесь связана с переходом от распределения $G(y)$ к распределению $F(x)$ [4]. В связи с этим будем считать, что вид ФР $F(x)$ нам неизвестен и задача относится к классу непараметрических задач оценки параметра сдвига.

В настоящее время нет недостатка в робастных оценках параметра сдвига, что создает даже определенное неудобство для пользователей (см. например, [3; 5] и библиографические списки к ним). Отметим ряд особенностей таких оценок. Большинство из них робастны на классе и имеют низкую эффективность в отсутствии выбросов. Как выход предложены адаптивные оценки: в основном используется адаптация по параметру усечения, но не по виду $F(x)$ [3], или адаптация ведется по виду распределения $F(x)$, но функция и параметр усечения подбираются эвристически [6]. Эта работа Р. Берана интересна в двух аспектах: в ней, очевидно, впервые введены робастные непараметрические оценки плотности, а также использован метод подстановки на основе этих оценок для получения оценки параметра. Становится понятным, что робастные эффективные оценки должны быть адаптивными как по виду основного распределения, так и по отбраковке выбросов.

В данной статье на основе взвешенного метода максимального правдоподобия (ВММП) [7; 8] синтезированы адаптивные робастные непараметрические оценки и показано их использование для оценки линейных функционалов.

Взвешенный метод максимального правдоподобия. Пусть $F(x, \theta)$ – унимодальное непрерывное распределение с плотностью $f(x, \theta)$ и неизвестным параметром θ – принадлежит к классу унимодальных распределений и $x_1, ..., x_N$ – выборка НОР из распределения $F(x, \theta)$. Обозначим через $F_N(x)$ эмпирическую функцию распределения (ЭФР), а через $g(x, \theta)$ – априорную плотность распределения.

M-оценки неизвестного параметра θ можно определить на основе решения эмпирического уравнения вида

$$\int \varphi(x, \theta_N)dF_N(x) = 0, \tag{1}$$

где $\varphi(x, \theta)$ – оценочная функция.

Анализ критерия радикальности и алгоритмов устойчивых оценок [5] позволяет сделать вывод, что все эти оценки можно получить на основе ВММП с оценочной функцией $\varphi(x, \theta)$ вида

$$\varphi(x, \theta) = \left[\frac{\partial}{\partial\theta} \ln g(x, \theta) + \beta\right] g^l(x, \theta), \tag{2}$$

где l – параметр радикальности оценки; β – параметр, который определяется по условию несмещенности оценки, в нашем случае $\beta = 0$ [7].

Нетрудно заметить, что (2) определяет ВММП с весами $g^l(x, \theta)$. При $l = 0$ мы получаем оценки максимального правдоподобия (ОМП), при $l = 0,5$ – радикальные оценки, при $l = 1$ – оценки максимальной устойчивости (ОМУ) [5]. Физически роль параметра l вполне понятна и сводится к определению степени мягкого усечения как для удаленных выбросов, так и по форме априорного распределения. Таким образом, варьируя параметром l, можно получать эффективные оценки при локальных отклонениях распределения $F(x, \theta)$ от априорного в классе устойчивых оценок.

В непараметрическом случае, когда вид $g(x,\theta)$ неизвестен, заменим $g(x,\theta)$ в (2) непараметрической симметризованной оценкой Розенблатта–Парзена

$$g_N(x,\theta)=\frac{1}{h_N}\int K\left(\frac{2\theta-x-t}{h_N}\right)dF_N(t). \qquad (3)$$

Например, для нормального ядра уравнения для оценки параметров сдвига θ и масштаба λ принимают следующий вид [7; 8]:

$$\begin{cases}\dfrac{1}{N(N-1)}\sum\limits_{i=1\neq}^{N}\sum\limits_{j=1}^{N}(\theta_N-z_{ij})\cdot W_1(z_{ij})=0,\\ \dfrac{1}{N(N-1)}\sum\limits_{i=1\neq}^{N}\sum\limits_{j=1}^{N}\left[\left(\dfrac{\theta_N-z_{ij}}{\lambda_N}\right)^2-\dfrac{1}{l+1}\right]\cdot W_1(z_{ij})=0,\end{cases} \qquad (4)$$

где

$$W_1(z_{ij})=\exp\left\{-\frac{(\theta_N-z_{ij})^2}{\lambda_N^2}\right\}\times$$

$$\times\left[\frac{1}{N-1}\sum_{i\neq m=1}^{N}\exp\left\{-\frac{(\theta_N-z_{im})^2}{\lambda_N^2}\right\}\right]^{l-1};$$

$z_{ij}=\dfrac{x_i+x_j}{2}$ – полусуммы Уолша.

Рассмотрим обобщенную *M*-оценку θ_N параметра θ, которая определяется на основе решения эмпирического уравнения вида

$$\int\varphi(x,\theta_N,\vec{T}_N(x,\theta_N))dF_N(x)=0,$$

где $\vec{T}=(T_1,\ldots,T_k)^T$; $T_i=\int S_i(x,t,\theta)dF(t)$; $T_{iN}=\int S_i(x,t,\theta)dF_N(t)$.

В связи с ограниченностью объема статьи приведем без доказательства ряд результатов в окончательном виде.

Имеет место следующее представление:

$$\theta_N-\theta=\left[\int\frac{\partial}{\partial\theta}\varphi(x,\theta,\vec{T})dF(x)\right]^{-1}\cdot\int\psi(t,\theta)dF(t),$$

$$\psi(t,\theta)=\varphi(t,\theta,\vec{T}(t,\theta))+$$

$$+\sum_{i=1}^{k}\int S_i(x,t,\theta)\frac{\partial}{\partial T_i}\varphi(t,\theta,\vec{T}(t,\theta))dF(x).$$

При выполнении ряда ограничений $\sqrt{N}(\theta_N-\theta)$ имеет асимптотически нормальное распределение с дисперсией

$$\sigma^2=\left[\int\frac{\partial}{\partial\theta}\varphi(x,\theta,\vec{T})dF(x)\right]^{-2}\cdot\int\psi^2(t,\theta)dF(t). \qquad (5)$$

Техника доказательства основана на работах Г. М. Кошкина ([9]) и результаты имеют место для стационарных процессов со слабой зависимостью.

В параметрическом случае ($S_i=0$)

$$\varphi(x,\theta)=\left[\frac{\partial}{\partial\theta}g(x,\theta)\right]g^{l-1}(x,\theta). \qquad (6)$$

Выражение (5) определяет дисперсию параметрического ВММП (классические *M*-оценки) и при $l=0$ (5) совпадает с выражением для дисперсии ОМП, а при $l=1$ – с выражением для дисперсии ОМУ [7].

Для непараметрического ВММП

$$\varphi(x,\theta,T_1,T_2)=T_1(x,\theta)\cdot T_2^{l-1}(x,\theta),$$

$$S_1(x,t,\theta)=\frac{1}{h_N}K\left(\frac{2\theta-x-t}{h_N}\right),$$

$$S_2(x,t,\theta)=\frac{\partial}{\partial\theta}S_1(x,t,\theta).$$

Выражение (5) определяет дисперсию непараметрического ВММП в зависимости от *l*.

Зависимости дисперсии параметрической (рис. 1) и вариации непараметрической (типа «складного ножа» jackknife) (рис. 2) оценок ВММП для модели Тьюки с асимметричным засорением от параметра радикальности l ($0\le l\le 1$) приведены ниже (кривая *1* на рис. 1 – без выбросов, кривая *2* – 3 % выбросов, среднее – 4, кривая *3* – 10 % выбросов, среднее – 4; кривая 1 – на рис. 2 – без выбросов, кривая *2* – выброс – 5, кривая *3* – выброс – 11, $N = 39 + 1$ выброс).

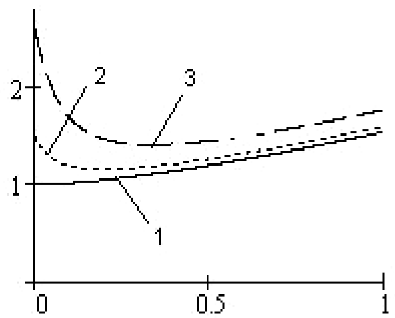

Рис. 1

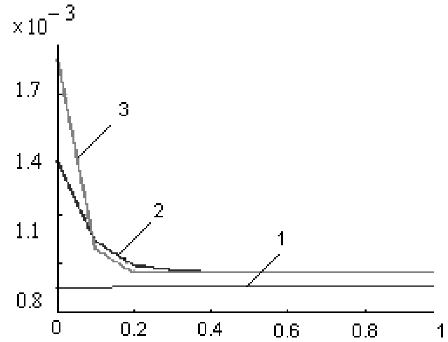

Рис. 2

Анализ дисперсии и вариации в зависимости от *l* (рис. 1, 2) показывает, что существует оптимальное *l*, доставляющее минимум дисперсии и вариации оценки.

Адаптивные оценки взвешенного метода максимального правдоподобного. Непараметрический подход на основе оценок Розенблатта–Парзена вида (3) позволяет осуществить адаптацию оценок ВММП по виду распределения. Адаптации по параметру радикальности l ($0 \le l \le 1$) производится с помощью бутстреп-метода. Для этого достаточно использовать простые бутстреп-процедуры типа «складного ножа» (jackknife) и алгоритмы поиска минимума вариации непараметрического ВММП. Моделирование также показывает, что при оптимальном l наблюдается и минимальное смещение оценки.

Примеры. Как отмечалось выше, значительный интерес представляет нахождение робастных непараметрических оценок для *U*-статистик. Применим для этого адаптивные оценки ВММП.

В первую очередь нас интересуют робастные непараметрические оценки функции распределения $G(t) = \int C(t-y)dG(y)$ и плотности в виде $g(t) = \int K((t-y)\cdot h_N^{-1})dG(y)$, где $C(y)$ – функция Хевисайда; $K(y)$ – ядерная функция. Зафиксируем значение $t = t_0$. От выборки $y_1, ..., y_M$ перейдем к выборкам $x_i = C(t_0 - y_i)$ для ФР и $x_i = K((t_0 - y_i)\cdot h_N^{-1})$ для плотности соответственно.

Представим результаты моделирования в зависимости от l для асимметричной модели выбросов Тьюки ($N = 100$, 10 % выбросов из нормального распределения со средним, равным пяти, рис. 3, 4). Хорошие результаты показывают радикальные оценки ($l = 0{,}5$), l оптимально при $l = 0{,}35$, при $l = 1$ происходит достаточно сильное подрезание.

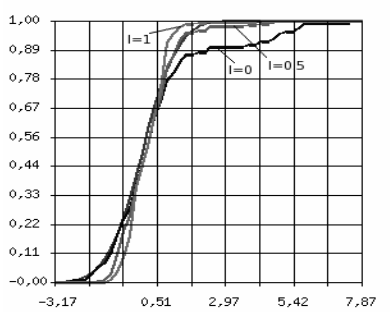

Рис. 3

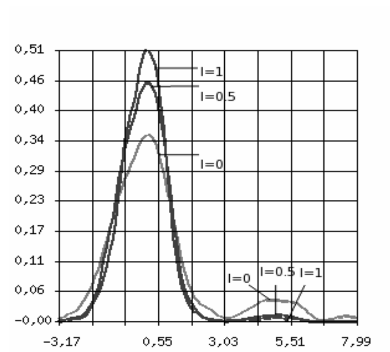

Рис. 4

Результаты моделирования для вариаций оценок дисперсии ($x_k = 0{,}5\cdot(y_i - y_j)$) и средней разницы Джини ($x_k = |y_i - y_j|$) приведены на рис. 5, 6 ($N = 30$ + + 1 выброс).

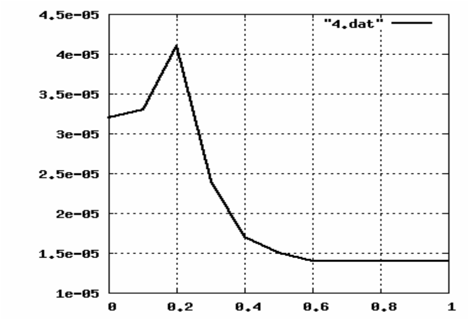

Рис. 5

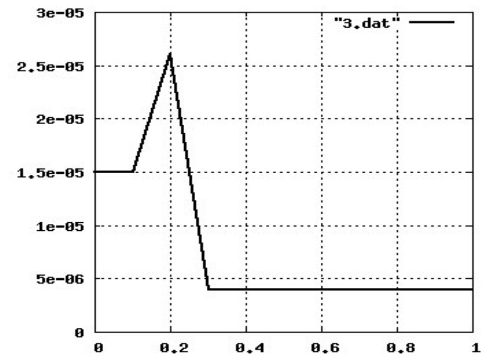

Рис. 6

Таким образом, предложен адаптивный робастный непараметрический алгоритм нахождения линейных функционалов, который позволяет адаптивно (путем мягкого усечения) настраивать оценку в зависимости от исходного распределения и выбросов. Рассмотрено робастное оценивание функции распределения, плотности распределения типа Розенблатта–Парзена, дисперсии, средней разницы Джини. Проведено моделирование оценок для асимметричной модели засорений Тьюки. На модели эксперимента Берана [7] проведено сравнение оценки Берана и вышеприведенной оценки. Они показывают одинаковые результаты, но в оценке Берана функция усечения и окно для нее (адаптация) подбирались эвристически [6]. Необходимо отметить, что представленный в данной статье подход позволяет применять робастные оценки ФР и плотности методом подстановки для получения адаптивных оценок неявных параметров от нелинейных функционалов.

Библиографические ссылки

1. Королюк В. С., Боровских Ю. В. Теория *U*-статистик. Киев : Наук. думка, 1989.

2. Непараметрическое оценивание функционалов по стационарным выборкам / Ю. Г. Дмитриев,

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.

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 предпосылкам. 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 характеристик, 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-