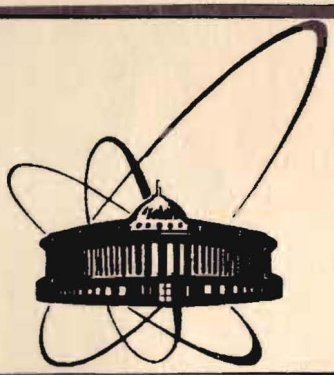


сообщения
объединенного
института
ядерных
исследований
дубна

Р5-85-492

А.А.Астапов, В.П.Бородюк*, С.В.Куняев*,
Г.А.Ососков, Н.И.Чернов

ЧИСЛЕННЫЙ АНАЛИЗ РОБАСТНЫХ РЕГРЕССИОННЫХ МЕТОДОВ

* Московский энергетический институт

1985

Введение

Многие проблемы, возникающие при автоматической обработке данных трековых камер в физике высоких энергий, принадлежат к кругу регрессионных задач математической статистики. В качестве актуального примера можно привести задачу выделения (фильтрации) точек, относящихся к изучаемому физическому событию, среди множества всех точек, поступивших в память ЭВМ при автоматической оцифровке изображения этого события в трековой камере. События состоят из треков – следов частиц, выходящих из вершины – точки взаимодействия. Оцифрованный трек – это дискретная последовательность точек $p_i, i=\overline{1,n}$, связанных близостью к гладкой кривой – проекции траектории частицы в магнитном поле. В основе алгоритмов фильтрации трековых данных лежит модель, предполагающая нормальное распределение $N(0, \sigma^2)$ для e_i – – отклонения точки p_i от этой кривой, что позволяет успешно применять для определения параметров треков хорошо разработанный аппарат метода наименьших квадратов (МНК), т.е. минимизировать функционал, составленный из квадратов указанных отклонений: $L=\sum e_i^2$.

Однако возможные пропуски оцифровок, появление различного рода шумовых отсчетов и, особенно, смешивание отсчетов от близко идущих и пересекающихся треков, что характерно для событий при современных высоких энергиях пучков, – все это приводит к засорению выборки, т.е. значимым нарушениям предположения о ее однородности и нормальности. Как известно (см., напр., /1/), в таких случаях МНК-оценки теряют свои оптимальные свойства. Взамен предлагаются так называемые робастные методы оценки /2,10/, состоящие в минимизации функционала $L_F=\sum F(e_i/\hat{\sigma})$ ($\hat{\sigma}$ – оценка σ) с функцией потерь F, растущей медленнее, чем квадратичная.

По определению П.Хьюбера, внесшего наибольший вклад в теорию таких оценок, робастность означает нечувствительность статистической процедуры к малым отклонениям от предположений принятой модели и, более того, бо́льшие из них не должны приводить к катастрофическим

последствиям (см. /2/, стр. I3, I4). Близким, но не эквивалентным является понятие устойчивости оценки (там же, стр. I6), которая означает нечувствительность к малым изменениям в основной выборке (имеются в виду малые изменения всех или большие изменения нескольких значений). Предложенная П.Хьюбером функция вклада

$$F(u) = \begin{cases} u^2/2 & , \ |u| \le c, \\ c|u| - c^2/2 & , \ |u| > c \end{cases} \qquad (I)$$

(c – параметр, зависящий от степени засорения) обеспечивает единственный минимум функционала L_F, состоятельность оценки и ее оптимальность в минимаксном смысле (см. /2/). Однако, как показали теоретические и практические исследования (см., например, /I4,4/), неограниченная функция вклада типа (I) не может обеспечить устойчивость оценок при наличии больших или асимметричных засорений. В таких случаях используют ограниченные функции вклада, краткий обзор которых дан в следующем разделе. К сожалению, неизбежным следствием отказа от выпуклости F явилась потеря единственности решения из-за возникновения дополнительных локальных минимумов функционала L_F. В работе /4/ было сделано предложение о повышении робастности оценок с невыпуклой функцией вклада с помощью специального вида шкалирования отклонения (стьюдентизации).

Настоящая работа посвящена качественному и количественному исследованию предложенного в /4/ метода путем численного анализа на модельных экспериментах, охватывающих различные способы засорения.

I. Робастные методы регрессионного анализа. Перейдем к строгим формулировкам, используя терминологию работ /4,II/. Рассмотрим регрессионную зависимость $y=\sum x_j b_j + e$, где x_j – факторы, определяемые местом, в котором производятся измерения; y – отклик (результат измерения); e – стационарная случайная ошибка; b_j – неизвестные регрессионные параметры, $j=\overline{1,m}$. Выборка наблюдений отклика при n различных значениях факторов записывается в матричном виде как

$$Y = XB + E \ , \qquad (2)$$

где $X=(x_{ij})$, $j=\overline{1,m}$, $i=\overline{1,n}$ – регрессионная матрица значений факторов.

Далее будет рассматриваться обобщенный, или взвешенный, метод наименьших квадратов (МНК) /II/, согласно которому оценка вектора параметров B записывается в виде

$$\hat{B} = (X^T W X)^{-1} X^T W Y \ . \qquad (3)$$

Здесь W – диагональная матрица $n \times n$, на диагонали которой расположены веса $w_i \ge 0$, приписываемые полученным значениям отклика, $i=\overline{1,n}$. Оценка вектора отклика определяется как $\hat{Y}=X\hat{B}$, а оценка вектора ошибок (так называемый вектор остатков) как

$$\hat{E} = Y - \hat{Y} = XB + E - X\hat{B} = (I - HW)E \ , \qquad (4)$$

где через H обозначена симметрическая $n \cdot n$-матрица

$$H = X(X^T W X)^{-1} X^T \ . \qquad (5)$$

В случае равноточных наблюдений, когда $\mathrm{cov} E = \sigma^2 I$, из (4)-(5) получаем

$$\mathrm{cov}\ \hat{E} = \sigma^2 (I - HW - WH + HW^2 H) \ . \qquad (6)$$

Пусть $x_k = \mathrm{col}(x_{k1}, \ldots, x_{km})$ – вектор значений факторов в k-м наблюдении отклика. Тогда k-й остаток $\hat{e}_k$ можно выразить как

$$\hat{e}_k = e_k - x_k^T (X^T W X)^{-1} X W E \ . \qquad (7)$$

Легко проверить, что $M\hat{e}_k = 0$, а дисперсия $D\hat{e}_k$ является k-м диагональным элементом матрицы (6), т.е. $D\hat{e}_k = \sigma^2 d_k^2$, где

$$d_k = (1 - 2w_k x_k^T (X^T W X)^{-1} x_k + x_k^T (X^T W X)^{-1} (X^T W^2 X) (X^T W X)^{-1} x_k)^{1/2} . \qquad (8)$$

Как отмечено во введении, МНК-оценки теряют свои оптимальные свойства при нарушении предположений о нормальности распределения ошибок e_i. Это связано с тем, что квадратичная функция вклада функционала L придает слишком большой вес наблюдениям с далекими отклонениями e_i, вследствие чего при тяжелых "хвостах" распределения e_i оценка $\hat{B}$ становится неустойчивой. В качестве примера распределения с тяжелыми "хвостами" Дж.Тьюки ввел модель ε-засоренного нормального распределения с параметром засорения ε $(0 \le \varepsilon \le 1)$:

$$f(x) = (1 - \varepsilon) N(x) + \varepsilon G(x) \ , \qquad (9)$$

где $N(x)$ – нормальное $N(0, \sigma^2)$, а $G(x)$ – произвольное засоряющее распределение.

Минимизация функционала L_F (см. введение) с функцией вклада F, растущей медленнее квадратичной, предложена многими авторами. Указанный во введении метод Хьюбера (I) при $c=0$ переходит в метод наименьших модулей (МНМ), исследованный в книге /5/. Заметим, что МНМ дает оценки максимального правдоподобия для ошибки с распределением Лапласа $f(e) = (2\lambda)^{-1} \exp(-|e|/\lambda)$. Отметим также метод Форсайта /6/ с функцией вклада $F(u) = |u|^p$, $0 < p < 2$. Следуя П.Хьюберу /2/, будем называть оценки, полученные минимизацией функционала L_F, М-оценками.

Для вычисления М-оценок в случае функций вклада общего вида используется итеративная МНК-процедура Флетчера – Гранта – Хеблена (ФГХ) /8/, в которой решается система нормальных уравнений, определяющих минимум L_F :

$$\sum_i \Psi\left(\frac{e_i}{\sigma}\right) x_{ij} = 0 \ ; \quad j=\overline{1,m} \ , \qquad (I0)$$

где $\Psi(u) = dF(u)/du$. На $(q+1)$-й итерации оценка $\hat{B}^{(q+1)}$ вычисляется по формуле (3), с определением весов $w_i^{(q+1)}$ через оценку $\hat{B}^{(q)}$ из формулы (4) на предыдущей итерации: $w_i^{(q+1)} = \Psi(\hat{e}_i^{(q)}/\hat{\sigma})/\hat{e}_i^{(q)}$. Если нет априорных начальных данных, то на первой итерации $\hat{B}^{(1)}$ может быть получена по МНК с единичной весовой матрицей W. Функцию $w(u) = \Psi(u)/u$ принято называть весовой функцией М-оценки.

последствиям (см. /2/, стр. I3, I4). Близким, но не эквивалентным является понятие устойчивости оценки (там же, стр. I6), которая означает нечувствительность к малым изменениям в основной выборке (имеются в виду малые изменения всех или большие изменения нескольких значений). Предложенная П.Хьюбером функция вклада

$$F(u) = \begin{cases} u^2/2 & , \ |u| \le c, \\ c|u| - c^2/2 & , \ |u| > c \end{cases} \qquad (I)$$

(c – параметр, зависящий от степени засорения) обеспечивает единственный минимум функционала L_F, состоятельность оценки и ее оптимальность в минимаксном смысле (см. /2/). Однако, как показали теоретические и практические исследования (см., например, /I4,4/), неограниченная функция вклада типа (I) не может обеспечить устойчивость оценок при наличии больших или асимметричных засорений. В таких случаях используют ограниченные функции вклада, краткий обзор которых дан в следующем разделе. К сожалению, неизбежным следствием отказа от выпуклости F явилась потеря единственности решения из-за возникновения дополнительных локальных минимумов функционала L_F. В работе /4/ было сделано предложение о повышении робастности оценок с невыпуклой функцией вклада с помощью специального вида шкалирования отклонения (стьюдентизации).

Настоящая работа посвящена качественному и количественному исследованию предложенного в /4/ метода путем численного анализа на модельных экспериментах, охватывающих различные способы засорения.

I. Робастные методы регрессионного анализа. Перейдем к строгим формулировкам, используя терминологию работ /4,II/. Рассмотрим регрессионную зависимость $y=\sum x_j b_j + e$, где x_j – факторы, определяемые местом, в котором производятся измерения; y – отклик (результат измерения); e – стационарная случайная ошибка; b_j – неизвестные регрессионные параметры, $j=\overline{1,m}$. Выборка n наблюдений отклика при различных значениях факторов записывается в матричном виде как

$$Y = XB + E \ , \qquad (2)$$

где $X=(x_{ij})$, $j=\overline{1,m}$, $i=\overline{1,n}$ – регрессионная матрица значений факторов.

Далее будет рассматриваться обобщенный, или взвешенный, метод наименьших квадратов (МНК) /II/, согласно которому оценка вектора параметров B записывается в виде

$$\hat{B} = (X^T W X)^{-1} X^T W Y \ . \qquad (3)$$

Здесь W – диагональная матрица $n \times n$, на диагонали которой расположены веса $w_i \ge 0$, приписываемые полученным значениям отклика, $i=\overline{1,n}$. Оценка вектора отклика определяется как $\hat{Y}=X\hat{B}$, а оценка вектора ошибок (так называемый вектор остатков) как

$$\hat{E} = Y - \hat{Y} = XB + E - X\hat{B} = (I - HW)E \ , \qquad (4)$$

где через H обозначена симметрическая $n \cdot n$-матрица

$$H = X(X^T W X)^{-1} X^T \ . \qquad (5)$$

В случае равноточных наблюдений, когда $\mathrm{cov}E = \sigma^2 I$, из (4)–(5) получаем

$$\mathrm{cov}\,\hat{E} = \sigma^2 (I - HW - WH + HW^2 H) \ . \qquad (6)$$

Пусть $X_k = \mathrm{col}(x_{k1}, \ldots, x_{km})$ – вектор значений факторов в k-м наблюдении отклика. Тогда k-й остаток $\hat{e}_k$ можно выразить как

$$\hat{e}_k = e_k - X_k^T (X^T W X)^{-1} X W E \ . \qquad (7)$$

Легко проверить, что $M\hat{e}_k = 0$, а дисперсия $D\hat{e}_k$ является k-м диагональным элементом матрицы (6), т.е. $D\hat{e}_k = \sigma^2 d_k^2$, где

$$d_k = (1 - 2w_k X_k^T (X^T W X)^{-1} X_k + X_k^T (X^T W X)^{-1} (X^T W^2 X) (X^T W X)^{-1} X_k)^{1/2} . \qquad (8)$$

Как отмечено во введении, МНК-оценки теряют свои оптимальные свойства при нарушении предположений о нормальности распределения ошибок e_i. Это связано с тем, что квадратичная функция вклада функционала L придает слишком большой вес наблюдениям с далекими отклонениями e_i, вследствие чего при тяжелых "хвостах" распределения e_i оценка $\hat{B}$ становится неустойчивой. В качестве примера распределения с тяжелыми "хвостами" Дж.Тьюки ввел модель ε-засоренного нормального распределения с параметром засорения ε $(0 \le \varepsilon \le 1)$:

$$f(x) = (1 - \varepsilon) N(x) + \varepsilon G(x) \ , \qquad (9)$$

где $N(x)$ – нормальное $N(0, \sigma^2)$, а $G(x)$ – произвольное засоряющее распределение.

Минимизация функционала L_F (см. введение) с функцией вклада F, растущей медленнее квадратичной, предложена многими авторами. Указанный во введении метод Хьюбера (I) при $c=0$ переходит в метод наименьших модулей (МНМ), исследованный в книге /5/. Заметим, что МНМ дает оценки максимального правдоподобия для ошибки с распределением Лапласа $f(e) = (2\lambda)^{-1} \exp(-|e|/\lambda)$. Отметим также метод Форсайта /6/ с функцией вклада $F(u) = |u|^p$, $0 < p < 2$. Следуя П.Хьюберу /2/, будем называть оценки, полученные минимизацией функционала L_F, M-оценками.

Для вычисления M-оценок в случае функций вклада общего вида используется итеративная МНК-процедура Флетчера – Гранта – Хеблена (ФГХ) /8/, в которой решается система нормальных уравнений, определяющих минимум L_F :

$$\sum_i \Psi\left(\frac{e_i}{\sigma}\right) x_{ij} = 0 \ ; \quad j=\overline{1,m} \ , \qquad (I0)$$

где $\Psi(u) = dF(u)/du$. На $(q+1)$-й итерации оценка $\hat{B}^{(q+1)}$ вычисляется по формуле (3), с определением весов $w_i^{(q+1)}$ через оценку $\hat{E}^{(q)}$ из формулы (4) на предыдущей итерации: $w_i^{(q+1)} = \Psi(\hat{e}_i^{(q)}/\hat{\sigma})/\hat{e}_i^{(q)}$. Если нет априорных начальных данных, то на первой итерации $\hat{B}^{(1)}$ может быть получена по МНК с единичной весовой матрицей W. Функцию $w(u) = \Psi(u)/u$ принято называть весовой функцией M-оценки.

Описанные выше варианты М-оценок используют неограниченную функцию вклада Ψ. Как отмечалось выше, такие оценки не могут быть устойчивыми в смысле Хьюбера, т.к. варьирование лишь одного наблюдения может привести к неограниченному смещению оценки $\hat{B}$. Хотя эти оценки и являются робастными, но их робастные свойства выражены относительно слабо (см.п.3, также [14]). Поэтому, если принятая модель допускает большие (и тем более неограниченно большие) ошибки наблюдений с высокой вероятностью появления, то более обоснованным становится применение М-оценок с ограниченной функцией вклада.

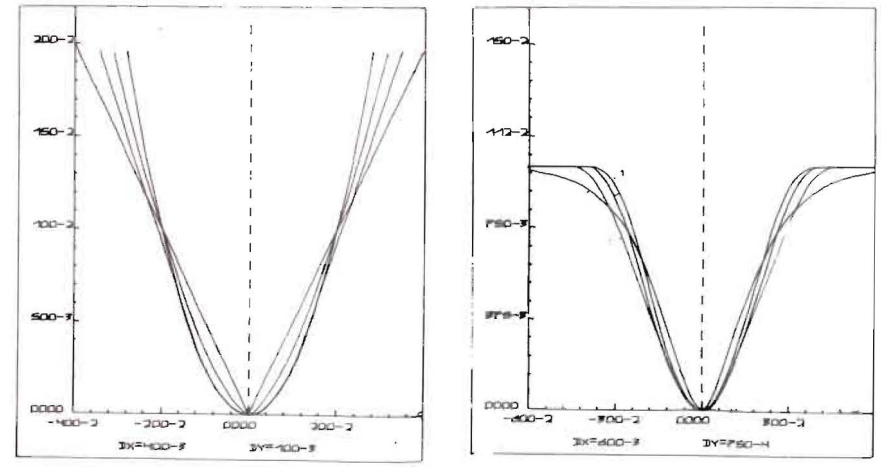
Предложен целый ряд таких оценок, различающихся конкретным видом функции вклада. Отметим методы Андрюса [7], Рамсея с экспоненциально убывающим весом [9], Тьюки с биквадратной весовой функцией(бивесом) [10], Хэмпела с кусочно-линейной производной функции вклада [13].

Таблица I

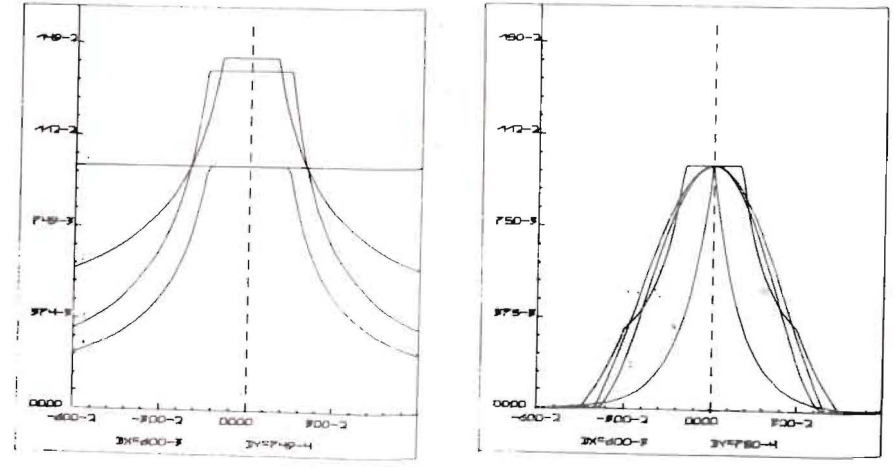
Метод	Весовая функции
Андрюса	$w(u)=\begin{cases}\frac{\sin(u/c)}{u/c}, & \lvert u\rvert < c\pi \\ 0, & \lvert u\rvert \geq c\pi\end{cases}$
Рамсея	$w(u)=\exp(-\lvert u/c\rvert)$
Тьюки	$w(u)=\begin{cases}(1-(u/c)^2)^2, & \lvert u\rvert < c \\ 0, & \lvert u\rvert \geq c\end{cases}$
Хэмпела	$w(u)=\begin{cases}1, & \lvert u\rvert \leq c_1 \\ c_1/\lvert u\rvert, & c_1 < \lvert u\rvert \leq c_2 \\ \frac{c_1(c_3-\lvert u\rvert)}{\lvert u\rvert(c_3-c_2)}, & c_2 < \lvert u\rvert < c_3 \\ 0, & \lvert u\rvert \geq c_3\end{cases}$
Д Р В	$w(u)=\frac{1+\lvert u/c\rvert^p\cdot(1-p/2)}{(1+\lvert u/c\rvert^p)^2}$

В таблице I приведены аналитические выражения для весовых функций этих методов. Графики неограниченных и ограниченных функций вклада с соответствующими весовыми функциями приведены на рис. I и 2 соответственно. Отметим практическое совпадение функций Тьюки и Андрюса после соответствующего масштабирования (кривые I и 2 на рис. 2б).

Нами предлагается следующее обобщение описанных М-оценок:



а Рис.I б



а б

Рис.2

функция вклада определяется выражением

$$F(u)=\frac{u^2}{1+|u/c|^p}\ , \qquad (II)$$

зависящим от двух параметров – масштаба (c) и формы (p).

Меняя параметр p ($0\le p\le 2$), можно получить функции, близкие к любой из изображенных на рис. I и 2. При $p=0$ они эквивалентны МНК. Для больших $|u|$ функция (II) приближается к функции $c^p u^{2-p}$, что соответствует функции Форсайта с $p^*=2-p$, но от последней она выгодно отличается близостью к обычной квадратичной функции МНК при малых $|u|$. При $p=2$ (II) дает ограниченную функцию вклада, близкую к функциям Тьюки и Андрюса, но в отличие от них лишь асимптотически приближающуюся к своему пределу. Это улучшает условие экстремума L_F, т.к. исключаются "овраги" с горизонтальным дном и горизонтальные плоские участки поверхности функционала L_F (см. рис. 7 и 8). В соответствии с видом выражения (II) при целых p будем называть введенную нами функцию дробно-рациональной функцией вклада (ДРВ-функцией). Ее весовая функция приведена в табл. I, а на рис. 3 изображен ее график при различных p.

2. Нормировка остатков с учетом индивидуальных дисперсий (стьюдентизация). Вышеописанные М-оценки допускают существенное уточнение, предложенное авторами в /4/ на случаи появления таких наблюдений отклика y_k, у которых велика не только ошибка $|e_k|$, но и значения факторов $(x_{k1},\dots,x_{km})$ в пространстве $\mathbb{R}^m$ лежат "в стороне" от остальных факторов (ср. рис. 4а и 4б для $m=2$). По терминологии Дж.Себера /II/, мы в определении функционала L_F переходим от шкалированных остатков $\hat e_k/\hat\sigma$ к их стьюдентизированным вариантам $\hat e_k/\hat\sigma d_k$, где d_k определяется, согласно (8), через дисперсию $D\hat e_k$-остатка, полученного на предыдущей итерации ФГХ-процедуры.

Цель этого метода состоит в противодействии влияния резко выпадающих наблюдений, для которых $d_k\approx 0$, за счет чего они "протягивают" линию регрессии $\hat y=x\hat\beta$ сильнее остальных точек. Такие точки П.Хьюбер назвал точками разбалансировки /2/. В то же время исключение точек разбалансировки путем разного рода процедур выброса (см., напр., /I0,I2/) может привести к потере информации и точности оценки $\hat\beta$.

Предлагаемое нами определение весов по формуле

$$w_k=\Psi(e_k/\hat\sigma d_k)/(e_k/\hat\sigma d_k)\ , \qquad (I2)$$

в которой e_k и d_k пересчитываются по (7) и (8) после каждой ФГХ-итерации, позволяет полностью исключить влияние точек разбалансировки при больших $|e_k|$ (см. пример в /4/) и не терять информацию при малых $|e_k|$.

Для наглядного качественного объяснения метода "стьюдентизации" остатков при $m=2$ сравним изображения функционала L_F и функционала, соответствующего системе весов (I2):

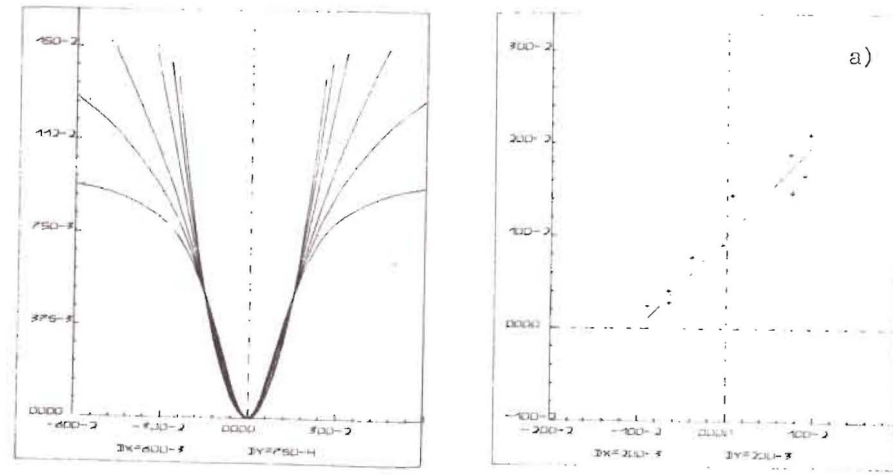


Рис.3

Рис.5

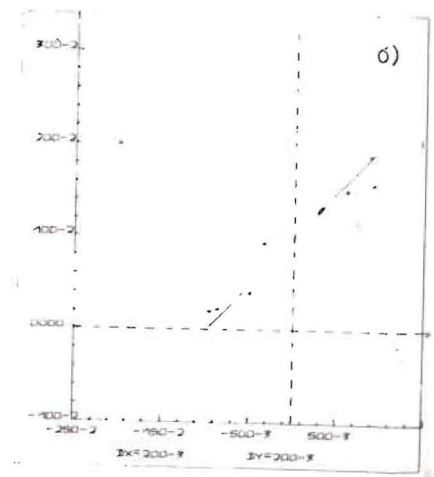
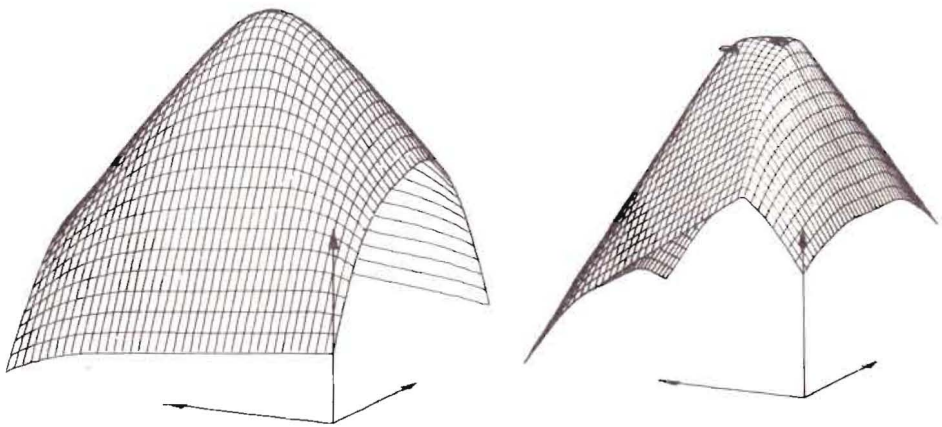


Рис.4

$$L_F^{(s)}(b_1,\ldots,b_m) = \sum_{i=1}^{n} d_i^2 F(e_i/\hat{\sigma}d_i) . \qquad (13)$$

Выражение $z=-L$ определяет гиперповерхность в $(m+1)$-мерном пространстве $(b_1,\ldots,b_m,z)$, точка $z(\hat{b}_1,\ldots,\hat{b}_m)=z_{max}$ соответствует искомой оценке $\hat{B}$.

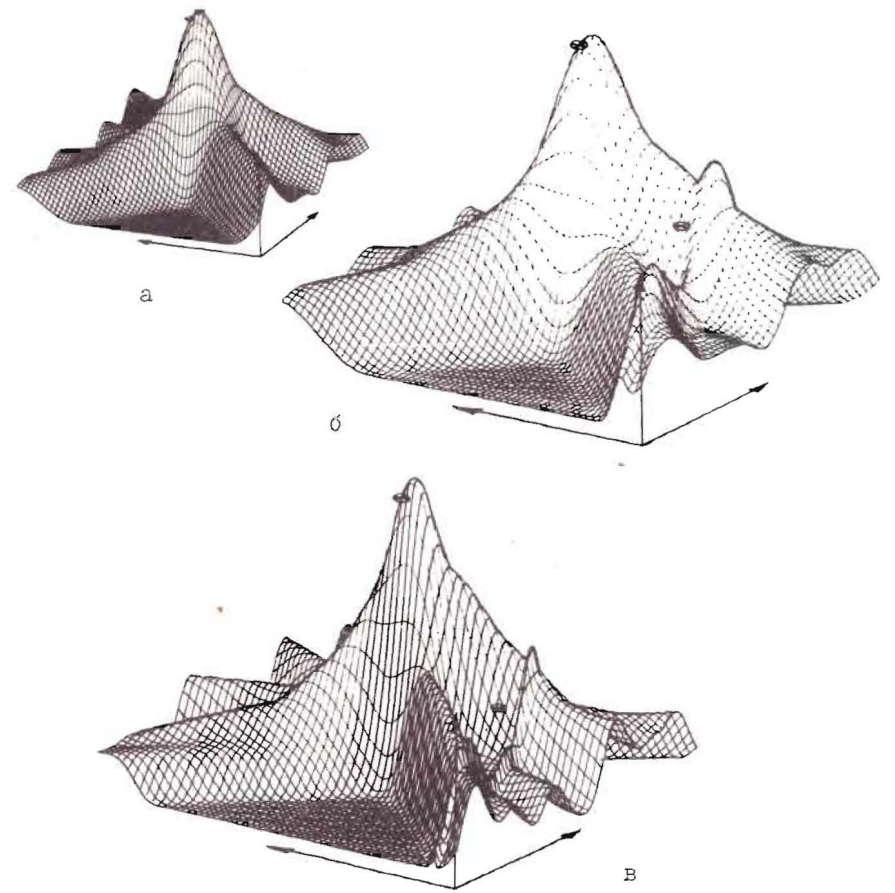
Для графического представления поверхности $z(b_1,b_2)$ была разработана программа, позволяющая изображать трехмерную поверхность на плоскости методом сечений с помощью графопостроителя, подсоединенного к ЭВМ СМ-4. На рис. 4а приведена зависимость $y=b_1x+b_2$, $b_1=b_2=1$, $-1\le x\le 1$ и точки выборки объема $n=10$, полученные добавлением к исходной зависимости нормально распределенной случайной ошибки с $\sigma=0{,}2$. На рис. 6а, 7а, 8а изображены поверхности $z=z(b_1,b_2)$, соответствующие функциям вклада Хьюбера, Тьюки, ДРВ $(p=2)$.



а б

Рис. 6

Появление точки разбалансировки при $x=-2$, изображенной на рис. 4б, резко изменило форму поверхностей функционалов. На рис. 5, 6б, 7б, 8б изображены поверхности для МНК, функций вклада Хубера, Тьюки и ДРВ $(p=2)$ соответственно. Хорошо видно уплощение поверхностей, соответствующих неограниченным функциям вклада, затрудняющее поиск экстремума. На рисунках отмечены точки, соответствующие первой ФГХ-итерации и истинные значения параметров $b_1=b_2=1$. На рис. 7б, 8б первая итерация оказывается вблизи локального экстремума, вызванного точкой разбалансировки, вследствие чего ФГХ-процедура часто приводит в этот "посторонний" экстремум. Стьюдентизация остатков (12)



а б в

Рис.7

вызывает резкое уменьшение площади горизонтальной проекции постороннего пика – см. на рис. 7в, 8в поверхности функционала $-L_F^{(s)}$ на первой ФГХ-итерации с функциями вклада Тьюки и ДРВ $(p=2)$. Благодаря этому уменьшается вероятность попадания первой итерации в область этого пика.

Необходимо исследовать также действие ФГХ-процедуры на последующих итерациях. Если вес $w_k^{(q)}$ точки – выброса p_k действительно будет уменьшаться с ростом номера итерации q, то одновременно будут расти и оценка остатка $\hat{e}_k^{(q)}$, и индивидуальная дисперсия $(d_k^{(q)})^2$. Существует ли опасность того, что таким образом уменьшение веса $w_k^{(q)}$ приведет к уменьшению отношения $\hat{e}_k^{(q)}/\hat{\sigma}^{(q)}d_k^{(q)}$ в (12) и тем самым

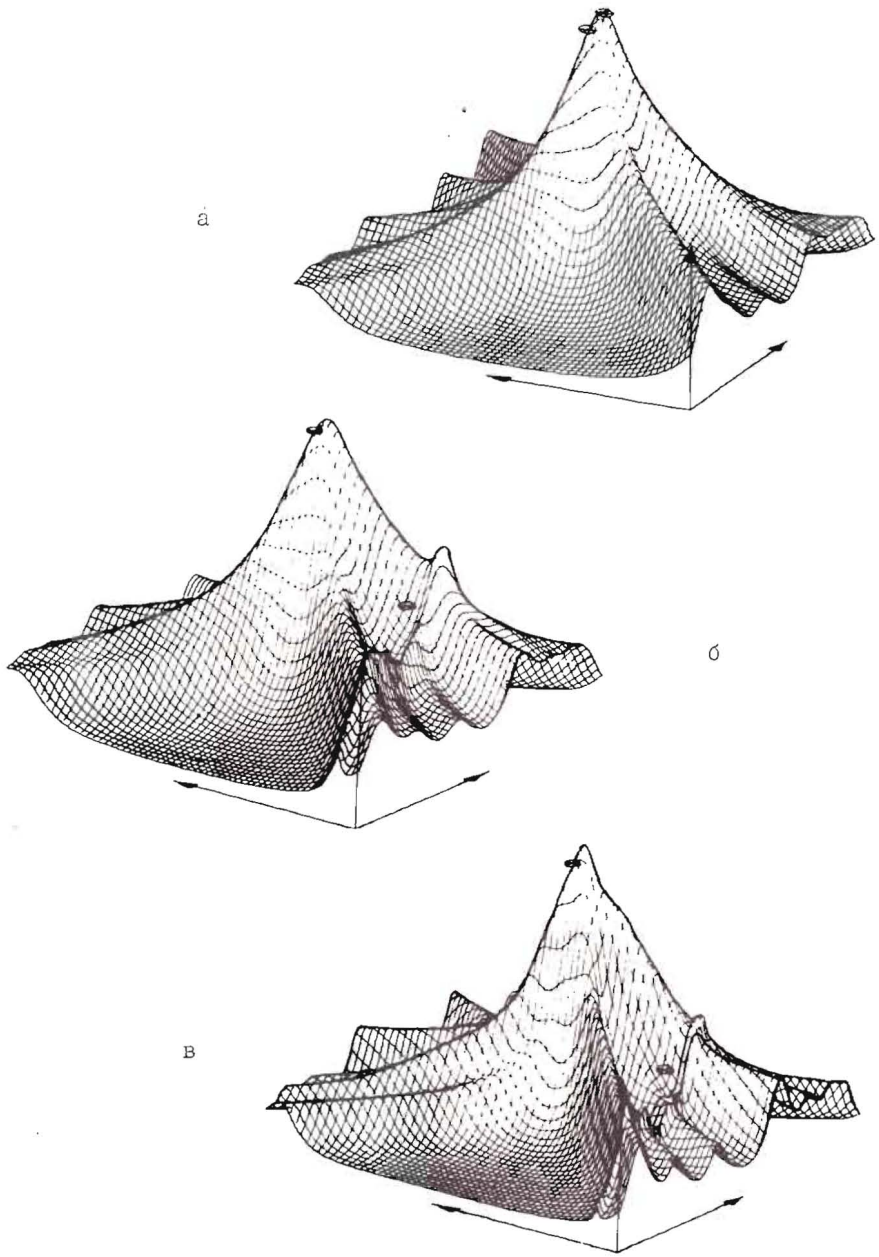


Рис 8

уже к увеличению веса $w_k^{(q+1)}$ на следующей итерации ("обратный ход" итерационной процедуры)? Как показывает следующая лемма, уменьшение веса $w_k^{(q)}$ любой точки не может непосредственно повлиять на вес $w_k^{(q+1)}$ этой же точки на следующей ТГХ-итерации.

Лемма. Величина $\hat{e}_k/d_k$ не зависит от веса к-й точки w_k.

Доказательство ведется в обозначениях п.I. Обозначим $f_k=1-x_k^T(x^TWx)^{-1}x_kW_k$ и $f_p=x_k^T(x^TWx)^{-1}x_pW_p$ при $p\neq k$. Тогда, в силу (7),

$$\hat{e}_k = \sum_{i=1}^{n} f_i e_i \quad . \tag{I4}$$

Достаточно доказать, что величины f_i/d_k не зависят от w_k при каждом $i=\overline{1,n}$. Заметим, что, в силу (I4),

$$\frac{f_i}{d_k} = \frac{f_i}{\sqrt{f_1^2+\ldots+f_n^2}} \quad . \tag{I5}$$

Обозначим $\alpha_k=1$ и $\alpha_p = x_k^T(\sum_{j\neq k} x_j^T w_j x_j)^{-1}x_pW_p$ при $p\neq k$ – величины, не зависящие от w_k. Имеем

$$\alpha_p f_k = (1-x_k^T(x^TWx)^{-1}x_kw_k)x_k^T(\sum_{i\neq k} x_i^Tw_ix_i)^{-1}x_pw_p = x_k^T\Big[(\sum_{i\neq k} x_i^Tw_ix_i)^{-1} -$$

$$-(x^TWx)^{-1}x_kw_kx_k^T(\sum_{i\neq k} x_i^Tw_ix_i)^{-1}\Big]x_pw_p = x_k^T(x^TWx)^{-1}\Big[(x^TWx)(\sum_{i\neq k} x_i^Tw_ix_i)^{-1} -$$

$$-x_kw_kx_k^T(\sum_{i\neq k} x_i^Tw_ix_i)^{-1}\Big]x_pw_p = x_k^T(x^TWx)^{-1}x_pw_p = f_p \quad ,$$

где мы воспользовались тем, что $x^TWx=\sum x_iw_ix_i^T$. Следовательно, $f_p=\alpha_pf_k$, откуда

$$\frac{f_i}{d_k} = \frac{\alpha_i}{\sqrt{\alpha_1^2+\ldots+\alpha_n^2}}\cdot\frac{f_k}{|f_k|} \quad .$$

Функция $f_k(w_k)$ непрерывна по w_k и поэтому, если $f_k(w_k)\neq 0$, то отношение f_i/d_k не зависит от w_k, что доказывает лемму. Если $f_k(w_k^{(o)})=0$ при некотором $w_k^{(o)}>0$, то $D\hat{e}_k=0$, что возможно лишь при специально подобранных матрицах x и W (см. /II/). Но в этом случае $\hat{e}_k(w_k)\equiv d_k(w_k)\equiv 0$. Лемма доказана.

3. Описание и результаты численного эксперимента. Специалисты по робастным оценкам отмечают трудности теоретических выводов в большинстве практических ситуаций, когда засорение выборки оказывается тяжелым и/или несимметричным и рекомендуют проверку предположений и методов на монте-карловских моделях.

Так, П.Хьюбер, отмечая /2/ "абсолютную беспомощность" асимптотической теории робастных оценок в практических случаях, предлагает уточнять результаты теории методом Монте-Карло, ссылаясь на свои же результаты /I3/. Большую практическую пользу рекомендаций, полученных путем статистического моделирования, отмечает и Дж.Тьюки /3/.

В соответствии с традициями работ /3,13/ для сравнения описанных в п.п. 1,2 методов был проведен численный эксперимент, в ходе которого моделировались различные способы засорения исходной регрессионной зависимости. Были рассмотрены два варианта исходной зависимости: двумерный $y=b_1x_1+b_2$ $(b_1=b_2=1)$ и пятимерный $y=b_1x_1+b_2x_2+b_3x_3+b_4x_4+b_5$ $(b_1=b_2=b_5=1, b_3=0.1, b_4=10)$. Случайный разброс точек генерировался по распределению (9), в котором нормальное распределение $N(x)$ с $\sigma=0,1$ засорялось нормальным распределением $G(x)$ со средним 0 и $\sigma=1$ (при $\varepsilon=0$, $\varepsilon=0,1$ и $\varepsilon=0,25$). Как и в /13/, отдельно исследован случай распределения Коши, когда $f(x)$ имеет плотность $10\pi^{-1}(1+(10x)^2)^{-1}$. Особое внимание было уделено моделированию тяжелого засорения, типичного для задач обработки трековой информации: А) наличие значительного разрыва в пространстве факторов $(\Delta x\sim 10)$, сопровождаемого значительным разбросом ($\sigma=5$) отклика в точке, лежащей после этого разрыва; засорение в виде менее плотного трека ($\varepsilon=0,25$): Б) пересекающего основной (среднее $y=-x$); В) идущего параллельно (среднее $y=1$). Дисперсия для случаев Б и В полагалась 0,01. Значения факторов x_i разыгрывались по нормальному закону со средним 0 и дисперсией 1, объем выборки полагался равным $N=10; 30; 100$.

Для подгонки регрессии по полученным точкам использовались 11 вариантов описанных в п.п. 1,2 методов: МНК, Форсайта ($p = 1,5$), Хьюбера ($c=1,4$), МНМ, Рамсея ($c=1$) , Андрюса ($c=1,15$) , Хэмпела ($c_1=1, c_2=2, c_3=3$) , Тьюки ($c=4$), ДРВ ($p=2, c=3$),а также варианты двух последних методов со стьюдентизацией остатков – п. 2.

Для оценки точности полученной подгонки $\hat{y}=\hat{Y}(x,\hat{b}_1,\ldots,\hat{b}_k)$ ($k=2$ или 5) вычислялось усредненное квадратичное отклонение от истинной регрессии $y^{(0)}=y^{(0)}(x,b_1,\ldots,b_k)$ в выбранных точках факторного пространства $d=N^{-1}\sum(\hat{y}_i-y_i^{(0)})^2$. Величина d вычислялась для 1000 вариантов случайного розыгрыша точек вокруг истинной регрессии с фиксированными значениями факторов x_i . В качестве меры точности исследуемых методов подгонки было выбрано среднее $\bar{d}=(d^{(1)}+\ldots+d^{(1000)})/1000$, значения которого, умноженные для удобства на 10^4 (для варианта А – на 10), для всех выбранных вариантов подгонки приведены в табл.2 и 3. Исследование временных затрат на расчет одной ФГХ-итерации при $k=2, N=10$ на ЭВМ ЕС-1060 показали, что вычисление индивидуальных дисперсий (8) занимает 1,0 мкс, вычисление весов w (табл. 1)– от 1,3 мкс до 2,4 мкс в зависимости от применяемого метода и на вычисление оценки (3) – 3,6 мкс. Необходимое число итераций для стабилизации трех значащих цифр параметров b_1 и b_2 колебалось от 3 до 6.Большое количество итераций (4–6) требовали методы Рамсея, Хэмпела и МНМ.

Данные таблиц 2 и 3 позволяют сделать следующие выводы:

1. В отсутствие засорения ($\varepsilon=0$) все робастные методы по точности уступают МНК, но не более чем на 10–15%.

2. При росте засорения до $\varepsilon=0,1$ МНК резко (на порядок) теряет в точности. Робастные методы показывают хорошую точность, кроме метода Форсайта, близкого к МНК.

3. При значительном засорении ($\varepsilon=0,25$) или при распределении Коши методы с неограниченной функцией вклада работают заметно хуже по сравнению с методами с ограниченной функцией вклада.

4. Стьюдентизация остатков при сравнительно небольшом выигрыше в точности (12–17% при $\varepsilon=0,25$ и до 30% для распределения Коши) дает очень хороший эффект при появлении точек разбалансировки (вариант А).

5. Модели А, Б, В, характерные для трековой информации, являются наименее подходящими для непосредственной оценки регрессионных параметров, т.к. слишком велика ошибка при начальном приближении, полученном при $w_i\equiv 1$. В этих ситуациях необходимо использовать априорную информацию, которую в практике распознавания треков (см., напр., /12/) получают путем экстраполяции с участков трека, уже прослеженных вне области засорения – см. также /4/.

В случае А процедуру оценки можно сделать двухэтапной: сначала проводится предварительная оценка по выборке с исключенными точками разбалансировки, а на втором этапе эта оценка используется как начальное приближение для робастной оценки по полной выборке. Как показали расчеты, это дает снижение квадратичного отклонения $\bar{d}$ на два порядка. Временные потери на стьюдентизацию остатков составляют до 20% (см. выше), поэтому имеет смысл вводить процедуру стьюдентизации в виде отдельной ветви, включаемой при появлении точек разбалансировки.

В заключение отметим, что специальный расчет для модели А,в котором начальное приближение для ФГХ-процедуры искусственно задавалось в точке постороннего локального минимума функционала L (см. рис. 7в, 8в) показал, что стьюдентизированная ФГХ-процедура тем не менее выходила к истинному (глобальному) минимуму в 22% просчитанных вариантов.

Авторы благодарят Г.И.Симонову за ряд полезных обсуждений и замечаний, И.Н.Силина за внимание к работе и ценные советы.

Таблица 2

	Метод	k=2, N=10							k=2, N=30			
		$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=0.25$	Коши	В	Б	А	$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=0.25$	Коши
I	МНК	20,I	204	5I5	$5\cdot10^5$	986	II50	2I3	6,4I	64,3	I6I	$3\cdot10^5$
2	Форсайт	2I,0	98,7	294	10^5	758	986	224	6,78	I5,8	45,7	I250
3	Хьюбер	2I,I	78,4	286	10^3	840	I020	2I8	6,7I	I0,2	32,5	32,I
4	МНМ	24,0	66,2	202	2I8	608	83I	239	7,45	I2,7	24,9	22,6
5	Рамсей	26,7	50,8	I68	I45	6I2	749	24I	8,54	9,45	I3,3	I8,0
6	Андрюс	23,I	63,I	230	339	750	9I7	223	7,39	8,53	I6,9	22,0
7	Тьюки	23,I	63,3	228	336	748	9I2	223	7,27	8,5I	I6,9	22,0
8	Тьюки (с)	23,5	50,7	20I	222	725	784	I35				
9	Хэмпел	25,2	60,6	204	252	687	825	229	8,I5	9,23	I5,2	20,8
I0	ДРВ	22,5	54,3	2II	328	70I	859	224	7,47	8,72	I6,3	2I,3
II	ДРВ (с)	22,4	52,0	200	237	694	800	I88				

Таблица 3

	k=2, M=100				k=5, N=30				k=5, N=100			
	$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=0.25$	Коши	$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=0.25$	Коши	$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=0.25$	Коши
I	I,9I	22,2	5I,5	$4\cdot10^5$	I6,3	I73	398	$I7\cdot10^5$	5,00	55,7	I28	$3\cdot10^5$
2	2,07	4,34	9,94	46,0	I7,2	56,8	I64	$9\cdot10^4$	5,I6	II,7	28,6	7I2
3	2,07	2,97	5,93	8,I2	I7,2	3I,I	II3	II2	5,24	7,49	I5,5	I9,3
4	2,30	4,04	5,58	5,87	I9,8	38,5	67,I	I46	5,47	6,65	20,3	84,5
5	2,44	3,08	3,69	4,70	24,0	29,8	44,0	60,0	6,35	7,I2	8,93	II,8
6	2,II	2,48	3,49	5,78	I9,6	25,2	63,3	68,6	5,50	6,33	8,83	I4,0
7	2,II	2,49	3,49	5,75	I9,6	25,2	62,0	68,2	5,52	6,3I	8,86	I4,0
8					I9,2	25,I	58,2	67,9				
9	2,40	2,66	3,35	5,42	22,2	28,I	50,9	67,0	6,I0	6,85	8,7I	I3,5
I0	2,24	2,55	3,57	5,64	I8,7	24,9	59,4	64,I	5,70	6,52	9,26	I3,4
II					I8,4	24,7	58,5	64,0				

ЛИТЕРАТУРА

I. Айвазян С.А. и др. Прикладная статистика. Финансы и статистика, М., I985.

2. Хьюбер П. Робастность в статистике. Мир, М., I984.

3. Tukey J.W. in Robustness in Statistics . N.-Y.Acad. Press, 1979.

4. Кунаев С.В. и др. ОИЯИ, Р10-84-553, Дубна, I984.

5. Мудров В.И., Кушко В.Л. Метод наименьших модулей. Знание, М., I97I.

6. Forsythe A.B. Technometrics, 1972 vol. 14,No.2,p.159-166.

7. Andrews D.F. Technometrics, 1974,vol. 16, No.4,p.523-531.

8. Fletcher R. et al. Comput. J.,1971, vol. 14, No.3, p.276.

9. Ramsay J.O., J.Amer. Statist.Assoc., 1977, vol.72,No.3,p.608-615.

I0. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Финансы и статистика, М., I982.

II. Себер Дж. Линейный регрессионный анализ. Мир, М., I980.

I2. Ососков Г.А. ОИЯИ, Р10-83-I87, Дубна, I983.

I3. Huber P.J. Ann. Math. Statist. 1972, v. 43, No.4, p.1041-1067.

I4. Щурыгин А.М. В кн.: П Всесоюзная школа - семинар "Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа". М., I983, с. 73-8I.

Рукопись поступила в издательский отдел
25 июня I985 года

Вниманию организаций и лиц, заинтересованных в получении публикаций Объединенного института ядерных исследований

Принимается подписка на препринты и сообщения Объединенного института ядерных исследований.

Установлена следующая стоимость подписки на 12 месяцев на издания ОИЯИ, включая пересылку, по отдельным тематическим категориям:

ИНДЕКС	ТЕМАТИКА	Цена подписки на год
1.	Экспериментальная физика высоких энергий	10 р. 80 коп.
2.	Теоретическая физика высоких энергий	17 р. 80 коп.
3.	Экспериментальная нейтронная физика	4 р. 80 коп.
4.	Теоретическая физика низких энергий	8 р. 80 коп.
5.	Математика	4 р. 80 коп.
6.	Ядерная спектроскопия и радиохимия	4 р. 80 коп.
7.	Физика тяжелых ионов	2 р. 85 коп.
8.	Криогеника	2 р. 85 коп.
9.	Ускорители	7 р. 80 коп.
10.	Автоматизация обработки экспериментальных данных	7 р. 80 коп.
11.	Вычислительная математика и техника	6 р. 80 коп.
12.	Химия	1 р. 70 коп.
13.	Техника физического эксперимента	8 р. 80 коп.
14.	Исследования твердых тел и жидкостей ядерными методами	1 р. 70 коп.
15.	Экспериментальная физика ядерных реакций при низких энергиях	1 р. 50 коп.
16.	Дозиметрия и физика защиты	1 р. 90 коп.
17.	Теория конденсированного состояния	6 р. 80 коп.
18.	Использование результатов и методов фундаментальных физических исследований в смежных областях науки и техники	2 р. 35 коп.
19.	Биофизика	1 р. 20 коп.

Подписка может быть оформлена с любого месяца текущего года.

По всем вопросам оформления подписки следует обращаться в издательский отдел ОИЯИ по адресу: 101000 Москва, Главпочтамт, п/я 79.

Астапов А.А. и др. P5-85-492

Численный анализ робастных регрессионных методов

Работа посвящена робастным оценкам параметров линейной регрессии. Цель работы состоит в численном анализе и модификации данных оценок для моделей, близких к задачам распознавания треков частиц . Для случая появления точек разбалансировки предложен специальный метод стьюдентизации вектора остатков и проведено теоретическое и качественно-графическое исследование этого метода. Результаты монте-карловских оценок численных характеристик робастных методов для различных моделей сведены в таблицу. Приведен ряд практических рекомендаций для использования робастных регрессионных методов в массовой обработке экспериментальных данных.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ

Сообщение Объединенного института ядерных исследований. Дубна 1985

Перевод О.С.Виноградовой

Astapov A.A. et al P5-85-492

Numerical Analysis of Robust Regression Methods

The robust estimates of linear regression parameters are considered. Numerical analysis of these methods and their modifications for the problems of particle track recongition have been performed. For the cases when disbalance points can appear a special method of studentizing of residual vector is proposed. This method is investigated by theoretical and graphical means. The numerical characteristics of the methods obtained by the Monte-Carlo simulation in various models are summarized in tables.Some recommendations for using robust regression methods in the mass data processing are given.

The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR

Communication of the Joint Institute for Nuclear Research. Dubna 1985