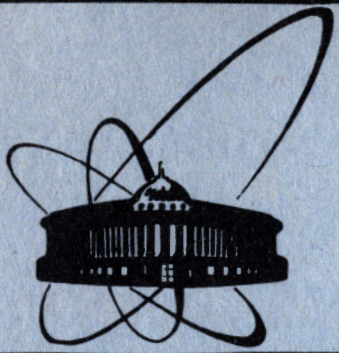

объединенный
институт
ядерных
исследований
дубна

Р10-84-553

С.В.Куняев,* Г.А.Ососков, Н.И.Чернов

СТАТИСТИЧЕСКИЕ МЕТОДЫ РАСПОЗНАВАНИЯ И РОБАСТНОГО ОЦЕНИВАНИЯ ПАРАМЕТРОВ ТРЕКОВ

Направлено на III Всесоюзный семинар
по обработке физической информации.
Цахкадзор, 9-12 октября 1984 г.

* Московский энергетический институт

1984

Настоящая работа посвящена исследованию методов статистического анализа, применяемых при обработке данных с трековых камер. Используемые алгоритмы реконструкции событий фактически приводят к задачам оценки параметров по выборке из неоднородной совокупности. В настоящее время этот раздел математической статистики интенсивно развивается благодаря богатым приложениям в физике, технике, экономике и т.д. (см.[1,2]). Результатом этого развития являются устойчивые алгоритмы высокой точности и надежности (так называемые робастные методы - от английского слова robust - крепкий, сильный), предложенные П.Хубером, Дж.Тьюки, Д.Андрюсом и др. Наша цель - модификация этих алгоритмов с учетом специфики обрабатываемой информации в физике высоких энергий.

Структура работы следующая. В первом параграфе описана модель камерного снимка и ставятся две задачи на применение статистических методов. Во втором параграфе дан краткий обзор робастных методов регрессионного анализа. В третьем предлагается модификация этих методов и показывается ее преимущество. Четвертый параграф посвящен усовершенствованию алгоритмов слежения по треку. В пятом приводится структура программы фильтрации данных сканирования снимков со стримерных камер на автомате АЭЛТ-2/I60 в ЛВТА ОИЯИ. В Приложение отнесен вывод формулы дисперсии точечных оценок при подгонке дугой окружности.

§ I. Данные автоматического сканирования изображения событий в трековой камере представляют собой сложную картину, состоящую из большого (10^4-10^5) числа отсчетов, расположенных вдоль нескольких треков (следов заряженных частиц) или порожденных шумовыми образованиями и сбоями сканирующего прибора. Это - типичная выборка из неоднородной совокупности. Возникающая задача оценивания параметров треков (число треков также случайно) весьма сложна как в теоретическом плане, так и для практических расчетов. Заметим, что не существует даже удовлетворительной математической модели такого набора экспериментальных данных (см.по этому поводу[3]). Известно лишь, что проекция трека на плоскость снимка может быть хорошо аппроксимирована

дугами окружностей. Уравнение дуги окружности $y=b \pm \sqrt{R^2-(x-a)^2}$ задает нелинейную (по параметрам a,R) регрессионную зависимость y от x. Необходимость использования нелинейного регрессионного анализа создает значительные трудности при расчетах на ЭВМ.

Оценивание параметров по выборке из неоднородной совокупности проводится в два этапа (см.[2]). Первый (задача № I) - оптимальное разделение неоднородной совокупности на однородные (в нашем случае - на отсчеты, принадлежащие каждому треку и на посторонние отсчеты). Второй (задача № 2) - оценивание параметров по каждой из выделенных групп. Реальное разделение на группы всегда содержит ошибки, поэтому приходится оценивать параметры по "засоренным" данным (отсчетам других треков и шумовым отсчетам). Уровень шумов достигает 30-50%. Обычные методы оценивания здесь уже не работают, а применяются специальные робастные методы, устойчивые к наличию помех и грубых ошибок (см.ниже).

Задача № I возникает, в частности, при обработке без предварительных целеуказаний оператора или при наведении по одной точке, когда возникает необходимость в отслеживании трека, проходящего через отмеченную оператором точку (например, вершину события - см.[17]). При "наведении по дорожке", построенной по трем точкам М, отмеченным оператором на треке (маске трека), обработка начинается сразу со второго этапа (задача №2).

§ 2. Для решения задачи № I существует широкий класс т.н. методов кластеризации - см.[4,5]. Их применение всегда основано на логическом анализе конкретной задачи, и вопрос об оптимальности, как правило, не решается. Один из таких методов мы предлагаем в § 3. Напротив, для решения задачи № 2 разработаны достаточно универсальные методы. Опишем их при довольно общей постановке задачи.

Рассмотрим регрессионную зависимость $y=\sum_j x_j b_j+e$, где x_j - факторы, y - отклик, e - стационарная случайная ошибка с нулевым средним и дисперсией σ^2, b_j - неизвестные регрессионные коэффициенты (параметры), $j=1,..,m$. Пусть имеется выборка значений отклика объема n, полученная в результате n независимых экспериментов при различных значениях факторов:

$$X=\begin{pmatrix} x_{11},\dots,x_{1m} \\ \vdots \\ x_{n1},\dots,x_{nm} \end{pmatrix}, \qquad Y=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Тогда справедливо

$$Y = XB + E \quad , \qquad (I)$$

где $B=(b_1,\dots,b_m)^T$, $E=(e_1,\dots,e_n)^T$ - неизвестные векторы. Случайный вектор E имеет скалярную ковариационную матрицу $covE=\sigma^2 I_n$.

В том случае, когда ошибка ε распределена по нормальному закону, оптимальной оценкой вектора параметров B является оценка по методу наименьших квадратов (МНК)[6]:

$$L(b_1,\dots,b_m)=\sum_i (y_i-\sum_j x_{ij}b_j)^2=\sum_i \hat{\varepsilon}_i^2 \to \inf \quad .$$

МНК - оценка $\hat{B}$ выражается через y линейно: $\hat{B}=(X^TX)^{-1}X^TY$.

В случае данных, "засоренных" выборками из совокупностей с другими законами распределения, плотность вероятности ошибки ε обычно имеет более тяжелые по сравнению с нормальным распределением хвосты. Известно, что в этом случае МНК-оценка не является удовлетворительной[2,7]. Вместо нее применяются специальные робастные методы оценивания. Большинство из них заключается в минимизации функционала

$$L_\psi(b_1,\dots,b_m)=\sum_i \psi\left(\frac{\hat{\varepsilon}_i}{\hat{\sigma}}\right) \quad , \qquad (2)$$

где $\psi(\cdot)$- надлежащим образом подобранная функция, а $\hat{\sigma}$ - оценка параметра масштаба (или $\hat{\sigma}=\sigma$, если величина σ известна).

Например, А.Форсайт[8] предлагает функцию $\psi(t)=|t|^\rho$ $\rho<2$. При $\rho=1$ метод Форсайта переходит в т.н. метод наименьших модулей, изучавшийся в [9,10]. П.Хубер [11] разработал алгоритм, использующий функцию

$$\psi(t)=\begin{cases} \frac{1}{2}t^2 & \text{при} \quad |t|<c \ ; \\ c|t|-\frac{1}{2}|t|^2 & \text{при} \quad |t|\geqslant c \ . \end{cases}$$

Метод Хубера корректен (функционал L_ψ имеет единственный локальный минимум) и оптимален в некотором минимаксном смысле - см.[12]. Однако оказалось, что этот метод работает только в "не слишком плохих" случаях. А именно, удовлетворительные оценки получаются лишь при не очень тяжелых хвостах плотности вероятности ошибки (коэффициент эксцесса не более 5÷6). Для случаев с большей засоренностью данных используются методы с невыпуклой функцией ψ .

Например, Д.Андрюс[13] предложил функцию

$$\psi(t)=\begin{cases} -1-\cos(t/c) & \text{при} \quad |t|<c\pi \ ; \\ 0 & \text{при} \quad |t|\geqslant c\pi \ . \end{cases}$$

Гораздо удобнее для вычислений на ЭВМ метод Дж.Тьюки [7] с функцией

$$\psi(t)=\begin{cases} t^2(1-(t/c)^2)^2 & \text{при} \quad |t|<c \ ; \\ 0 & \text{при} \quad |t|\geqslant c \ . \end{cases}$$

Значение c рекомендуется выбирать в диапазоне от 4 до 6.

Реализуются описанные методы (кроме метода Андрюса) в виде итерационных процедур, где на каждом шаге применяется МНК с весами $w_i=\psi(\hat{\varepsilon}_i/\hat{\sigma})/\hat{\varepsilon}_i^2$, которые вычисляются через оценки $\hat{b}_1,\dots,\hat{b}_k$ на предыдущей итерации. В качестве параметра масштаба $\hat{\sigma}$ предлагается какая-либо робастная оценка величины σ , например, $\hat{\sigma}=\mathrm{med}\{|\hat{\varepsilon}_i|\}/0{,}6745$, где $\mathrm{med}\{a_i\}$ - медиана (средний член вариационного) ряда выборки $\{a_i\}$. Коэффициент 0,6745 подобран так, чтобы оценка $\hat{\sigma}$ была

несмещенной в случае нормального распределения. Заметим, что такая оценка хороша только при не очень сильном засорении (уровень шумов до 10%). При большем засорении коэффициент 1/0,6745 следует уменьшать.

§ 3. Методы с невыпуклой функцией ψ (Андрюса, Тьюки) дают гораздо лучшие результаты в случаях "тяжелого засорения", но имеют свои недостатки (например, неединственность локального минимума функционала L_ψ). Кроме того, в тех случаях, когда точки $\bar{x}_i=(x_{i1},\dots,x_{im})$ расположены в пространстве факторов неравномерно, отдельно лежащие точки могут оказывать решающее влияние на значения регрессионных коэффициентов и приводить к грубым ошибкам (см. пунктир на рис.1).

Для выявления таких точек применяется метод "складного ножа"[7], заключающийся в последовательном отбрасывании каждой точки и проведении линии регрессии через оставшиеся (например, по МНК). Если полученная линия прошла далеко от этой точки, то точка признается "выбросом" и исключается из набора данных. Однако этот метод весьма трудоемкий и далеко не всегда приводит к успеху.

Предлагаемая нами модификация описанных робастных методов позволяет в значительной степени преодолеть эти трудности. Суть модификации заключается в учете расположения точек выборки в пространстве факторов. Для этого вместо функционала (2) предлагается минимизировать функционал

$$L_\psi^{(o)}(b_1,\dots,b_m)=\sum_i \psi\left(\frac{\hat{\varepsilon}_i}{d_i}\right) \quad ,$$

где d_i^2 - дисперсия величины $\hat{\varepsilon}_i$ (т.н. дисперсия точечной оценки - см.[6]).

Иначе говоря, мы уточняем нормировочный коэффициент $1/\hat{\sigma}$ в аргументе функции ψ и добиваемся того, чтобы все величины $\hat{\varepsilon}_i/d_i$ (т.н. "стьюдентизированные остатки"[6]) имели нулевое среднее и единичную дисперсию независимо от расположения точки $\bar{x}_i$ в пространстве факторов. При этом губительное влияние отдельно лежащих факторов весьма эффективно исключается.

Для иллюстрации рассмотрим пример на рис.1.

Через точки A_1÷A_7 требуется провести линию регрессии $y=ax+b$, точка A_7 - явный "выброс". В таблице 1 приведены результаты трех последовательных итераций при реализации метода Тьюки (константа $c=5$, оценка $\hat{\sigma}=\mathrm{med}\{|\hat{\varepsilon}_i|\}$). В таблице 2 приведены результаты трех последовательных итераций по методу Тьюки с описанной модификацией. Из таблиц видно, что таким образом удалось преодолеть притяжение "чужой" точки A_7.

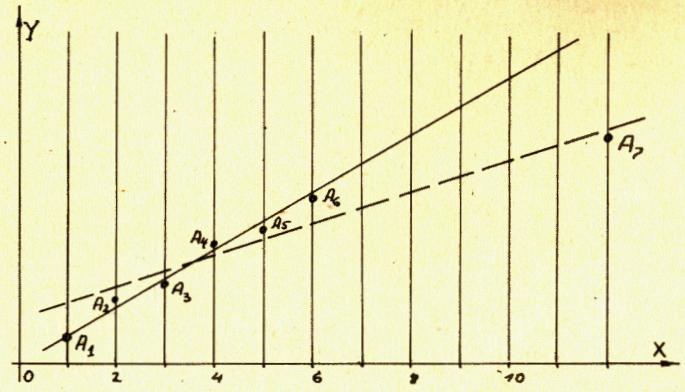

Рис.I

Таблица I

итерации			I		2		3	
i	x_i	y_i	$\hat{e}_i$	w_i	$\hat{e}_i$	w_i	$\hat{e}_i$	w_i
I	I.00	0.42	-0.48	0.68	-0.5I	0.59	-0.53	0.53
2	2.00	I.24	-0.04	I.00	-0.07	0.99	-0.08	0.99
3	3.00	I.67	0.00	I.00	-0.02	I.00	-0.03	I.00
4	4.00	2.29	0.23	0.92	0.2I	0.92	0.20	0.92
5	5.00	2.65	0.20	0.94	0.I9	0.94	0.I9	0.93
6	6.00	3.24	0.40	0.76	0.40	0.74	0.40	0.72
7	I2.00	4.85	-0.32	0.85	-0.29	0.86	-0.27	0.86

$\hat{\sigma}$ =0.23 | $\hat{a}$ =0.38 | $\hat{b}$ =0.5I → $\hat{\sigma}$ =0.2I | $\hat{a}$ =0.38 | $\hat{b}$ =0.54 → $\hat{\sigma}$ =0.20 | $\hat{a}$ =0.38 | $\hat{b}$ =0.56

Таблица 2

	I			2			3		
i	$\hat{e}_i$	d_i	w_i	$\hat{e}_i$	d_i	w_i	$\hat{e}_i$	d_i	w_i
I	-0.48	0.94	0.55	-0.47	0.82	0.46	-0.42	0.7I	0.42
2	-0.04	0.99	I.00	-0,05	0.7I	0.99	-0.03	0.58	I.00
3	0.00	I.03	I.00	-0.03	0.77	I.00	-0.03	0.65	I.00
4	0.23	I.05	0.9I	0.I8	0.8I	0.9I	0.I5	0.68	0.90
5	0.20	I.05	0.93	0.I3	0.78	0.94	0.08	0.64	0.97
6	0.40	I.04	0.72	0.3I	0.78	0.70	0.24	0.63	0.73
7	-0.32	0.49	0.35	-0.53	0.69	0.I8	-0.75	0.86	0.05
	$\hat{\sigma}$ = 0.23			$\hat{\sigma}$ =0.I8			$\hat{\sigma}$ = 0.I5		
	$\hat{a}$ = 0.38, $\hat{b}$ = 0.5I			$\hat{a}$ = 0.4I, $\hat{b}$ = 0.47			$\hat{a}$ =0.43, $\hat{b}$ =0.40		

Начиная с 4-й итерации, $w_7 \equiv 0$.

Численные эксперименты показывают, что модифицированные оценки значительно реже страдают от влияния далеко лежащих "выбросов".

Приведем формулу для вычисления дисперсии точечных оценок. Имеем

$$d_i^2 = \mathrm{Var}(y_i - \vec{x}_i\hat{B}) = \mathrm{Var}y_i + \mathrm{Var}(\vec{x}_i\hat{B}) - 2\mathrm{cov}(y_i, \vec{x}_i\hat{B}).$$

МНК-оценка $\hat{B}$ со взвешиванием выражается формулой $\hat{B}=(X^TWX)^{-1}X^TWY$, где W - диагональная матрица весов. Поэтому

$$d_i^2=\sigma^2+\vec{x}_i M((\hat{B}-B)(\hat{B}-B)^T\vec{x}_i^T)-2M(e_i\vec{x}_i(\hat{B}-B))=\sigma^2+\vec{x}_i(X^TWX)^{-1}X^TWM(EE^T)\times$$

$$\times((X^TWX)^{-1}X^TW)^T\vec{x}_i^T - 2M(e_iX_i(X^TWX)^{-1}X^TWE)=\sigma^2(1+H_iH_i^T-2H_iA_i),$$

где $H_i=\vec{x}_i(X^TWX)^{-1}X^TW$, а $A_i=(0,0,\ldots,0,1,0,\ldots,0)$ - i -й базисный вектор в R^n .

Отсюда $d_i=\sigma f_i$, где $f_i=\sqrt{1+H_iH_i^T-2H_iA_i}$ - величина, не зависящая от σ . В случае неизвестной величины σ можно, как и ранее, заменять ее оценкой $\hat{\sigma}$. Заметим, что при обработке камерных снимков параметр σ - средний разброс отсчетов вокруг трека - бывает приблизительно известен ($\sim$ четверть средней ширины трека). Практический опыт показывает, что известное приближенное значение σ (даже с ошибкой в I,5-2 раза) в случаях тяжелого засорения лучше, чем любая его оценка.

§ 4. Введение нормировочных коэффициентов типа дисперсий точечных оценок с использованием робастных процедур приводит к усовершенствованию одного алгоритма кластеризации для решения задачи № I (см. § I), который мы кратко опишем. Принцип работы этого алгоритма состо-

ит в объединении точек, лежащих в ряд на некоторой кривой в одну группу (кластер) с целью отнесения всех точек этой группы к одному из треков. Такие группы называются трек-элементами. Формируются они так называемым методом шнуров с экстраполяцией, т.е. путем последовательного поиска точек, при котором очередная точка ищется справа от уже найденных в некотором районе вокруг линии, фитирующей трек-элемент по уже найденным точкам (рис.2, подробности см. в [14]).

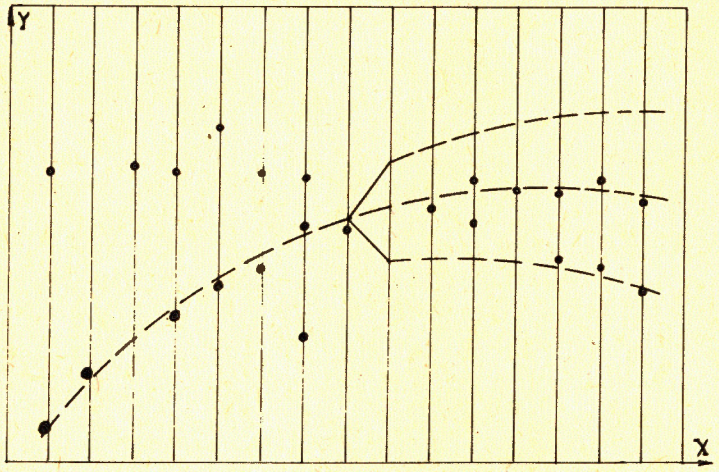

Рис.2

Если для фитирования использовать дугу окружности или аппроксимирующий ее полином, то в подобную процедуру фактически укладываются все известные методы прослеживания треков. Эти методы основаны на многократном применении МНК для отслеживания трек-элементов и для последующего выбрасывания посторонних точек, далеко отстоящих от фитирующей линии. Кроме трудоемкости, такая процедура страдала неустойчивостью, т.к. МНК не давал удовлетворительных оценок в условиях засоренности данных.

Использование робастных методов фитирования из § 3 значительно повышает точность экстраполяции и позволяет отказаться от громоздкой процедуры выброса, т.к. посторонним точкам автоматически приписываются нулевые веса. Наше предложение по повышению эффективности заключается в экстраполяции сразу целых сегментов отслеживаемого трека, т.е. присоединения к нему на каждом шаге не одной, а целой группы точек.

Проблема включения в трек-элемент нескольких точек с разными значениями фактора (координаты x на рис.2) решается следующим образом. Для каждой точки (x, y) из района поиска составляется статистика $n(x,y)=\rho(x,y)/d(x)$, где $\rho(x,y)$ – расстояние от точки (x, y) до линии экстраполяции, а $d^2(x)=\mathrm{Var}\,\rho(x,y)$ при условии, что точка (x, y) входит в данный трек.

Благодаря такой нормировке величина $n(x,y)$ имеет нулевое среднее и единичную дисперсию для каждой точки (x, y), и для включения ее в трек можно использовать постоянную пороговую процедуру $|n(x,y)|<K$ (или другой однородный критерий). Если трек-элемент фитируется по МНК прямой линией или полиномом, то линейная зависимость от параметра дает $d^2(x)=\sigma^2 g(x_1,\dots,x_n,x)$, где $x_1,\dots,x_n$ – абсциссы отслеженных ранее точек трек-элемента; функция g будет приведена ниже. Отметим полезный факт, что приближенная формула для $d^2(x)$ в случае нелинейного фитирования дугой окружности имеет такой же вид (см.Приложение).

В случае неизвестного σ вместо статистики $n(x,y)$ можно использовать величину $n_1(x,y)=\sigma^2 n^2(x,y)=\rho^2(x,y)\,g(x_1,\dots,x_n,x)^{-1}$ и пороговую процедуру $n_1(x,y)<K_1$.

При выводе явной формулы для функции g вернемся к общей регрессионной модели (1), описанной в § 2. Пусть $\vec{x}_0=(x_{01},\dots,x_{0m})$ – произвольная точка в пространстве факторов. Тогда

$$g(\vec{x}_1,\dots,\vec{x}_n,\vec{x}_0)=1+\sigma^{-2}\mathrm{Var}(\vec{x}_0\hat{B})=1+\|\vec{x}_0(X^TWX)X^TW\|^2 . \qquad (3)$$

Несмотря на кажущуюся громоздкость этой формулы, она выражает дробно-рациональную зависимость от x_{ij} и w_i и может быть включена в быстрый алгоритм отслеживания трек-элементов на ЭВМ.

Были проведены численные эксперименты с монте-карловской моделью двух пересекающихся треков. Вероятность обнаружения каждого трека на скан-линии была принята 0,64. Разброс точек вокруг оси трека разыгрывался по нормальному закону с параметрами $(0,\sigma^2)$. Среднее число шумовых точек на скан-линии было $\approx 0{,}8$. Рассматривалось два варианта модели: прямолинейные треки и треки в виде окружности.

В случае прямолинейных треков и фитирования прямой линией среднее время работы описанного выше алгоритма по сравнению с методом шнуров сокращается в 3–4 раза, причем точность оценок параметров трека возрастает.

Хорошие результаты получены и в случае криволинейных треков постоянной кривизны при фитировании дугой окружности. Этот способ требует бо́льшего начального набора точек, с которого начинается отсле-

живание (число точек не менее 10–15). Если такая начальная статистика набрана (методом поворотных гистограмм, шнуров или линейной экстраполяции), то дальнейшая процедура слежения успешно разделяет треки, пересекающиеся под малым углом.

Допустимая величина угла выражается через максимальную ширину трека h и длину L. Из рис.3 видно, что необходимое условие разделения треков есть $L > 2h/\sin\alpha \approx 12\sigma/\sin\alpha$ (мы принимаем $h \approx 6\sigma$). Для описанной процедуры с учетом начальной статистики можно считать достаточным условие

$$L > \frac{2h}{\sin\alpha} + L_0 \quad , \qquad (4)$$

где L_0 – минимальная длина участка трека, содержащего 10–15 точек.

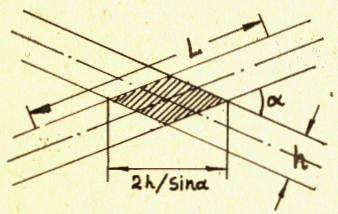

Рис.3

Описанные численные эксперименты показали, что вероятность сбоя процедуры (неправильного прослеживания трека) не зависит от угла α (проверялись значения α от $1{,}5^0$ до 20^0), а определяется числом N_0 точек на треке, отслеженных до области пересечения с другим треком (т.е. на участке длины L_0 в (4)). Зависимость вероятности сбоя P_0 от N_0 приведена в таблице 3.

Таблица 3

N_0	4	6	8	10	12	14	16	18	20
P_0	0,63	0,41	0,21	0,098	0,033	0,022	0,019	0,013	0,009

Для сравнения укажем, что метод шнуров с экстраполяцией полиномом второй степени (без процедуры выбрасывания посторонних точек) работает в 1,5 раза медленнее, а сбои дает в 2÷4 раза чаще.

§ 5. Изложенные выше методы оказались полезными при создании программы обработки фотоснимков со стримерной камеры, оцифрованных на сканирующем автомате с электронно-лучевой трубкой АЭЛТ-2/160 (см. /18/) в ЛВТА ОИЯИ. Разработанный алгоритм распознавания и фильтрации треков может работать как при наведении по одной точке трека, так и для более зашумленных треков при наведении по дорожке. В сочетании с развитыми средствами диалога автомата АЭЛТ-2/160 это обеспечивает достаточную гибкость при обработке снимков сложных событий.

Структура алгоритма следующая. В случае целеуказания в виде одной точки на треке обработка начинается по данным пилотного сканирования всего снимка с прослеживания трека от указанной точки методом, описанным в § 3. В результате из исходной совокупности отсчетов выделяется подгруппа близких к траектории частицы точек. В случае целеуказаний в виде дорожки такая группа состоит из всех точек, лежащих в указанной дорожке.

В обоих случаях выделенная совокупность отсчетов сильно засорена (доля посторонних точек 10–25%). Обработка этого материала предусматривает этапы фильтрации, подгонки кривой и выделения мастер-точек.

Первый этап состоит в подгонке дуги окружности по всем выделенным отсчетам робастным методом Тьюки с использованием быстрой МНК-процедуры, описанной в /15/. Область ширины $8\hat{\sigma}$ вокруг полученной дуги образует "узкую дорожку", за которой располагается значительная доля шумов (рис.4а).

Для лучшей фильтрации необходимо исключить из набора данных не только отсчеты, лежащие вне узкой дорожки, но и все отсчеты из областей пересечений треков, царапин и т.д. Для нахождения таких областей вся группа данных на этапе 2 переводится в полярные координаты с началом в центре найденной дуги окружности. В такой системе координат узкая дорожка становится прямоугольной полосой (рис.4б). Методом поворотных гистограмм отслеживаются все линии шумовых направлений, вдоль которых группируются отсчеты, лежащие вне узкой дорожки (это пересекающие треки, царапины и т.д.). Области пересечения этих направлений с данным треком исключаются из набора данных (заштрихованы на рис.4б).

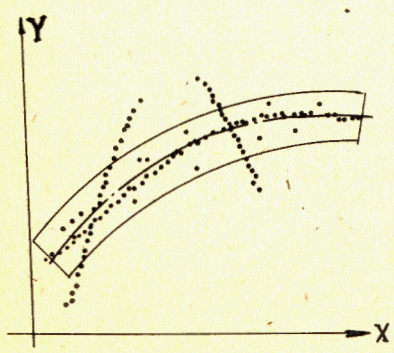

рис.4а

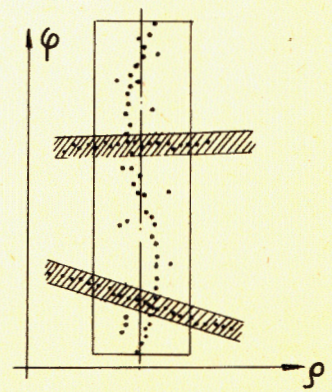

рис.4б

Отфильтрованные отсчеты не всегда хорошо аппроксимируются дугой окружности, т.к. треки часто имеют переменную кривизну (из-за искажений, неоднородности магнитного поля в камере и т.д. – см.рис.4а).

Поэтому в полярных координатах трек представляет собой S-образную кривую (рис.4б - см.также [16]).

На этапе 3 строится кусочно-циркулярная аппроксимация трека в декартовых координатах несколькими дугами окружностей разной кривизны, которая бы адекватно отражала поведение трека. Исходный трек делится пополам и каждая половина фитируется дугой окружности (см. [17]). Если аппроксимация не удовлетворительна, то эти участки снова делятся пополам и т.д. Если какой-либо из полученных участков не содержит достаточного количества отсчетов трека, то он исключается из рассмотрения.

Для проверки качества подгонки участка трека дугой окружности используется d-критерий Дурбина-Ватсона (см. [6], стр.164). Он состоит в следующем. Пусть $\hat{e}_i$, i=1,...,n - отклонения отсчетов данного участка трека от фитирующей дуги, упорядоченные вдоль трека. Тогда d -статистика

$$D=\frac{\hat{e}_1\hat{e}_2+\hat{e}_2\hat{e}_3+\ldots+\hat{e}_{n-1}\hat{e}_n}{\hat{e}_1^2+\hat{e}_2^2+\ldots+\hat{e}_n^2}$$

характеризует автокорреляцию остатков $\hat{e}_i$.

Аппроксимация считается удовлетворительной в случае $D<D_0$. Критический уровень D_0 подбирается экспериментально.

В заключение формируется массив мастер-точек трека. Они расставляются на каждой дуге из построенной кусочно-цикрулярной аппроксимации через равные промежутки друг от друга, но только в том случае, если в их окрестности присутствуют отсчеты данного трека.

Программа распознавания и фильтрации треков, разработанная на основе вышеописанных алгоритмов, была применена для обработки реальных данных сканирования 25 треков со стримерной камеры РИСК, полученных на автомате АЭЛТ-2/160 в режиме наведения по дорожке. Программа даже при отсутствии информации о предварительно измеренных точках в более чем 90% случаев позволяла отсеять участки пересечений треков, снизить уровень засоренности с 20-50% до 3-8% и выделить все точки, принадлежащие измеряемым трекам.

Программа показала хорошие результаты по скорости работы, надежности и возможности перенесения на малую машину класса СМ.

ПРИЛОЖЕНИЕ

Здесь мы выводим приближенную формулу для дисперсии точечной оценки $d^2(x)$ (см. § 4) в случае нелинейной регрессии, задаваемой уравнением дуги окружности $y=b\pm\sqrt{R^2-(x-a)^2}$. Для нахождения МНК-

оценок параметров a,b,R существуют лишь приближенные или итеративные методы. Мы используем неитеративный метод высокой точности, описанный в [15]. Предполагается, что исходные точки (x_i,y_i) лежат вдоль дуги, радиус которой значительно превышает ее длину (что вполне оправдано при отслеживании треков). В соответствующей системе координат уравнение дуги можно записать в виде $y=R-\sqrt{R^2-x^2}$ (рис.5). Наше предположение означает, что $\varepsilon=\ell/R$ (ℓ - длина дуги) - малый параметр и

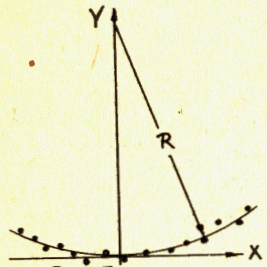

Рис.5

$$x_i=O(\varepsilon)R\ ,\quad \mathrm{Var}y_i=\sigma^2=R^2\cdot O(\varepsilon^4)\ , \qquad (5)$$

откуда

$$Ey_i=R-\sqrt{R^2-x_i^2}=\frac{x_i^2}{2R}+O(\varepsilon^4)R=O(\varepsilon^2)R\,. \qquad (6)$$

Пусть $x_0=O(\varepsilon)-R$ - некоторое произвольное значение фактора, тогда, как показано в § 4, $d^2(x_0)=\sigma^2+\mathrm{Var}(\hat{b}-\sqrt{\hat{R}^2-(x_0-\hat{a})^2})$.

Подставляя сюда явные выражения для $\hat{a}$, $\hat{b}$ и $\hat{R}$ из [15] и учитывая (5) и (6), получим

$$d^2(x_0)=\sigma^2 g_2(x_1,\ldots,x_n,x_0)+O(\varepsilon^6)R^2\ , \qquad (7)$$

где $\sigma^2 g_2(x_1,\ldots,x_n,x_0)$ - формула для дисперсии точечной оценки $d^2(x_0)$ при фитировании точек (x_i,y_i) полиномом второй степени (частный случай формулы (3) в § 4). Очевидно, $\sigma^2 g_2(x_1,\ldots,x_n,x_0)=O(\varepsilon^4)R^2$, поэтому, отбрасывая в (7) величину, на два порядка меньшую, получим искомую приближенную формулу дисперсии точечной оценки при фитировании дугой окружности:

$$d^2(x_0)\approx\sigma^2 g_2(x_1,\ldots,x_n,x_0)\ .$$

Относительная погрешность

$$P=\frac{d^2(x_0)-\sigma^2 g_2(x_1,\ldots,x_n,x_0)}{d^2(x_0)}$$

была оценена методом Монте-Карло. Проверялись значения $x_0>x_{(n)}=\max\{x_1,\ldots,x_n\}$, используемые в алгоритмах слежения с экстраполяцией в § 4. Зависимость p от величины экстраполируемой дуги $\Delta\varphi=(x_0-x_{(n)})/R$ изображена на графике (рис.6).

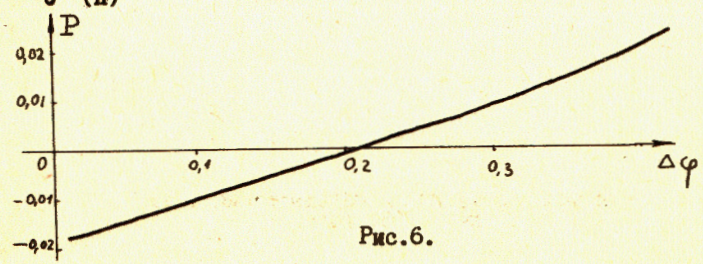

Рис.6.

Для алгоритма слежения из § 4 это вполне удовлетворительная погрешность.

Литература

1. Идье В. и др. Статистические методы в экспериментальной физике, перевод с англ. М., Атомиздат, 1976.
2. Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания, М., Статистика, 1980.
3. Бережной В.А. и др. Препринт ИВФЭ, 80-86, ОМВТ, 1980.
4. Классификация и кластер.(под ред. Дж. Вэн Райзина), М., Мир, 1980.
5. Дюран Б., Оделл П. Кластерный анализ. М., Статистика, 1977.
6. Себер Дж. Линейный регрессионный анализ. М., Мир, 1980.
7. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия, М., Финансы и статистика, 1982.
8. **Forsythe A.B. Technometrics, 1972, v. 14, p. 159-166.**
9. Мудров В.И., Кушко В.Л. Метод наименьших модулей, М., Знание,1971.
10. Колмогоров А.Н. Матем.сборник, 1931, т.38, с.47-50.
11. **Huber P.J. Annals of Math. Statistics, 1972, v.43, p.1041.**
12. Поляк Б.Т. В кн.: Структурная адаптация сложных систем управления. Воронеж, ВПИ, 1977, с.66.
13. **Andrews D.F. Technometrics, 1974,v.16, p. 523.**
14. Ососков Г.А. ОИЯИ, Р10-83-187, Дубна, 1983.
15. Ососков Г.А., Чернов Н.И. ОИЯИ, Р5-84-7, Дубна, 1984.
16. **Dunn L.A. in: Proc. of the Conference on Computing Assist. Scanning, Padova, 1976, p.75.**
17. Хофф П. В кн.: Автоматическая обработка данных с пузырьковых и искровых камер, М., Атомиздат, 1971, с.139.
18. Баранчук М.К.и др. ОИЯИ, Р10-88-61, Дубна, 1975.

Рукопись поступила в издательский отдел
30 июля 1984 года.

Куняев С.В., Ососков Г.А., Чернов Н.И.

Р10-84-553

Статистические методы распознавания и робастного оценивания параметров треков

Рассматриваются математические основы алгоритмов распознавания следов частиц при анализе данных с трековых камер. Подход к распознаванию как к задаче оценивания по выборке из неоднородной совокупности позволил применить робастные методы оценивания. Предложенное авторами усовершенствование этих методов позволило существенно увеличить точность и скорость процедуры прослеживания треков, что подтверждено на решении модельных задач. В заключение описаны алгоритмы фильтрации, основанные на предлагаемых методах, и их применение для обработки реальных данных сканирования снимков со стримерной камеры на автомате АЭЛТ-2/160.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1984

Перевод авторов