P13-2001-147

И.А.Голутвин, Ю.Т.Кирюшин, С.А.Мовчан,
Г.А.Ососков, В.В.Пальчик, Е.А.Тихоненко

РОБАСТНЫЕ ОПТИМАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ ТРЕК-СЕГМЕНТОВ МЮОНОВ В КАТОДНО-СТРИПОВЫХ КАМЕРАХ ЭКСПЕРИМЕНТА CMS

Направлено в журнал «Приборы и техника эксперимента»

1 Введение

Мюонная система является важной частью компактного мюонного соленоида (Compact Muon Solenoid, CMS) [1] - установки, которая создается на строящемся в ЦЕРН (Швейцария) крупнейшем в мире ускорителе — большом адронном коллайдере (Large Hadron Collider, LHC). Мюонная система CMS должна обеспечивать высокую эффективность восстановления мюонных треков с высоким пространственным разрешением в тяжелых фоновых условиях.

Эта система состоит из 4-х CSC(Cathode Strip Chambers)-камер, между которыми расположены стальные диски, изменяющие направление магнитного поля. CSC, катодно-стриповые (с нарезкой одного из катодов на полоски - радиальные стрипы) камеры, - это шестислойные многопроволочные пропорциональные камеры с дополнительным считыванием информации со стрипов.

Геометрия одного слоя и принцип работы CSC показаны на рис. 1. Заряженная

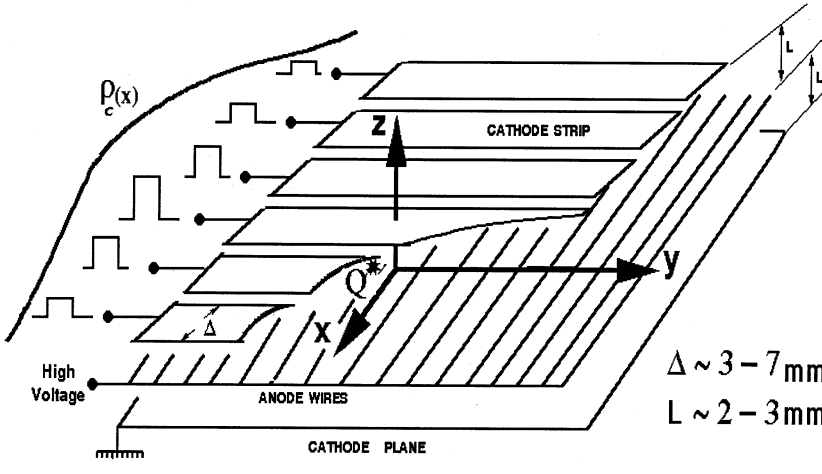

Рис. 1. Принцип работы катодной стриповой камеры

частица, проходя через газовый зазор между катодными плоскостями, ионизирует газ и вызывает формирование заряда $-Q$ на ближайшей анодной проволоке. В свою очередь, на катодных плоскостях индуцируется заряд $+Q$, который имеет некоторое известное пространственное распределение $\rho_c(x)$. Считывание этого заряда со стрипов катодной плоскости используется для точного определения координаты[1] частицы в азимутальном направлении (поперек стрипов).

[1]Т.е. вычисления центроида распределения зарядов, индуцированных на стрипах.

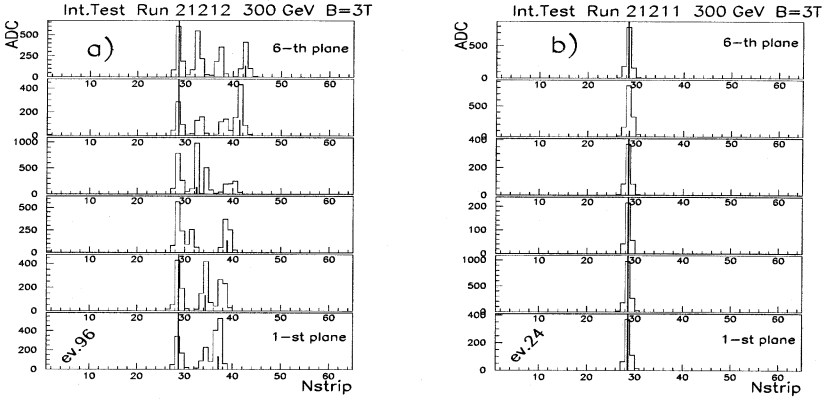

Рис. 2. Примеры мюонных событий на данных со стрипов, полученных на дубненском прототипе CSC. Заряд индуцируется на стрипах шести катодных плоскостей камеры. Восстановленные координаты центроидов распределения заряда показаны длинными отрезками: a) пример события, сильно загрязненного вторичными частицами; b) типичное событие с δ-электроном в 5-й плоскости, искажающим мюонный сигнал

Реконструкция трек-сегментов в CSC является одной из стадий восстановления траекторий μ-мезонов в установке CMS. Прежде чем реконструировать полный трек во всей мюонной системе, следует начать распознавание и фитирование трек-сегментов в каждой мюонной камере. Таким образом, необходимо найти оптимальный подход к фитированию трек-сегментов в условиях сильного загрязнения данных, протестировать его на модельных данных и использовать для обработки экспериментальных данных с прототипа.

Требуемое азимутальное пространственное разрешение для CMS мюонных CSC-камер составляет $\sim$ 100 мкм. От 10 до 30 % измеренных мюонных координат в CSC будут загрязнены различными источниками, но вторичное электромагнитное сопровождение (δ-электроны), порождаемое мюоном при прохождении вещества калориметра, расположенного перед CSC, и вещества самой камеры, является наиболее существенным при реконструкции координат и фитировании трека мюона в камере. Значительная часть δ-электронов проходит в непосредственной близости от траекторий мюонов, и во многих случаях невозможно разделить двойные срабатывания (разделение может быть осуществлено лишь в случаях, когда расстояние между заряженными частицами составляет несколько миллиметров). При высоких энергиях порядка сотен ГэВ число электронов сопровождения возрастает экспоненциально с ростом энергии. В результате загрязнения распределение ошибок измерений мюонных координат отличается от нормального (гауссова) распределения и имеет длинные негауссовы "хвосты". В таких случаях, как это хорошо

известно [2, 3], традиционный метод наименьших квадратов (МНК) теряет свои оптимальные свойства, в то время как робастные М-оценки [4] значительно менее чувствительны к загрязнению данных.

Пример события, сильно загрязненного вторичными частицами, представлен на рис. 2a. Типичное событие с δ-электроном, искажающим форму распределения заряда на стрипах, представлено на рис. 2b [2].

В работе [5] для CSC разработаны процедуры разделения перекрывающихся сигналов, распознавания, а также фитирования треков мюонов по МНК с последовательным удалением выбросных точек и повторным фитированием по оставшимся данным. Описанная в [5] процедура МНК с вычеркиванием выбросов и перефитированием [3] приводит к достаточно высокой эффективности восстановления треков, но оставляет открытым вопрос об эффективности оценок параметров треков. В работе [6] была сделана первоначальная попытка применить робастный подход для достоверной оценки параметров треков мюонов в CSC в условиях сильного загрязнения. Данная работа является логическим продолжением работ [5] и [6] и ставит своей целью решить проблему оптимального фитирования треков при наличии вторичных электронов, дающих срабатывания в непосредственной близости к мюонам.

Здесь и далее под понятием "оптимальные оценки" мы имеем в виду, в первую очередь, получение оценок параметров треков методом максимального правдоподобия (ММП) с минимально возможной дисперсией (такие оценки в математической статистике называются эффективными [7]). Для достижения этого в распределении ошибок измерений, входящих в функционал ММП, следует учитывать характер загрязнения данных. И второе предъявляемое нами требование к оптимальным оценкам состоит в том, что процедура вычисления этих оценок должна быть приближена по простоте к расчетам по МНК.

Таким образом, целями данной работы являются:

- создание математической модели CSC с учетом шума электроники и вторичных электромагнитных частиц, порождаемых мюоном при прохождении вещества;
- разработка и применение робастного подхода для оптимального фитирования трек-сегментов в CSC в условиях сильного загрязнения;
- сравнительный анализ параметров треков, полученных по МНК (с вычеркиванием выбросов и перефитированием) [5] и с помощью робастного метода;

[2] Кластер в 5-й плоскости (рис. 2b) можно рассматривать как пример загрязнения. Этот кластер сформирован зарядами, индуцированными несколькими заряженными частицами (μ и $\delta - e^-$). В результате центроид общего распределения заряда расположен в стороне от искомой мюонной координаты. Центроиды распределения зарядов, составляющих трек, обозначены на рисунке длинными отрезками, а лежащие в стороне от трека - короткими.

[3] Как подчеркивается в [10, с.102], удаление выбросов с последующим оцениванием уже является не чем иным, как робастными оценками. Поэтому процедуру МНК с вычеркиванием выбросов и перефитированием из [5] можно отнести к "слабым" робастным оценкам с неоптимальными весами [8].

- оценка азимутального пространственного разрешения в прототипе CSC наиболее достоверным образом.

2 Математическая модель и особенности реализации робастного подхода

Квадратичность функционала, минимизация которого осуществляется в МНК, приводит к тому, что далеко отстоящие точки могут дать неоправданно большой вклад в функционал и привести к значительной потере точности оценок параметров. Чтобы избежать этого, следует учитывать измерения только из непосредственной окрестности ($\sim [3 \div 4] \cdot \sigma$) подгоняемой функции, придавая остальным измерениям меньшие значения или вообще пренебрегая ими. Такую идею можно реализовать, придавая каждому измерению специальный вес, значение которого убывает с ростом расстояния до подгоняемой кривой. Устойчивый к выбросам подход, называемый **робастным**[4], был предложен П.Хьюбером [4]. Предложение Хьюбера сводится к некоторому обобщению метода максимального правдоподобия. Подчеркивая эту связь с ММП, Хьюбер назвал свой подход *М-оцениванием*. С математической точки зрения предлагалось перейти от суммы квадратов к сумме некоторых *функций вклада*, которые также зависят от отклонения d_i точки от регрессионной кривой, но растут медленнее, чем квадратичная парабола.

Итак, рассмотрим линейную регрессионную зависимость следующего вида:

$$x_i = \sum_{j=1}^{p} \phi_j(z_i) \cdot \theta_j \; + \; d_i \; , \quad i = 1, \ldots, N, \tag{1}$$

где
$\phi_j(z)$ - известный набор p линейно независимых геометрических функций (например, $1, \; z, \; \ldots, \; z^{p-1}$);
z_i - координата i-й катодной плоскости детектора;
x_i - "отклик" (измерение) в i-й плоскости детектора (в CSC это центроид распределения зарядов, измеренных на стрипах);
d_i - случайная ошибка измерения в этой плоскости детектора;
θ_j - неизвестные регрессионные параметры ($j = 1, \ldots, p$), которые следует оценить по выборке экспериментальных данных;
N - число плоскостей детектора.

Для загрязненного распределения ошибок измерения d_i используется так называемая модель "больших ошибок" (gross-error model) [4]:

$$f(e) = (1 - \epsilon) \cdot g(d) \; + \epsilon \cdot h(d), \tag{2}$$

являющаяся суперпозицией основного (g) и загрязняющего (h) распределений.

Здесь
g - распределение Гаусса $g(d_i) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{d_i^2}{2 \cdot \sigma^2}}$;

[4]Robust (англ.) — крепкий, здоровый. В статистике — не чувствительный к загрязнению данных.

$h(d_i) = \alpha \cdot e^{-\beta \cdot |d_i|}$ - двустороннее экспоненциальное распределение "больших ошибок" при загрязнении одним δ-электроном;
$\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 + \epsilon_5$ (ϵ_k - параметр суммы экспоненциальных загрязнений $h(d_i)$ не менее чем k δ-электронами) - суммарный параметр множественного загрязнения.

ϵ_k, α и β получены параметризацией результатов моделирования с помощью программы GEANT [9] и соответственно равны:
$\epsilon_1 = 0,23$, $\epsilon_2 = 0,07$, $\epsilon_3 = 0,02$, $\epsilon_4 = 0,006$, $\epsilon_5 = 0,002$,
$\alpha = 2,23^{-1}, \beta = 3,1^{-1}$.
Таким образом, $\epsilon = 0,328$.

Следует подчеркнуть, что согласно Хампелю [10] распределение, которое ведет себя в центральной части как нормальное, а на концах - как экспоненциальное, называется "наименее предпочтительным распределением Хьюбера"[5].

Далее, используя метод максимального правдоподобия ($L = \prod_{i=1}^{N} f(d_i) \longrightarrow max$), мы получаем систему уравнений

$$\sum_{i}^{N} w_i \cdot \phi_j(z_i) \cdot d_i \; + \; \beta \cdot \sigma^2 \cdot \sum_{i}^{N} \tilde{w_i} \cdot \phi_j(z_i) \cdot sign(d_i) \; = \; 0 \; , \; j = 1, \ldots, p, \qquad (3)$$

с оптимальными весами $w_i = \frac{1+c}{1+c \cdot e^{d_i^2/2 \cdot \sigma^2 - \beta \cdot |d_i|}}$ и

$$\tilde{w_i} \equiv 1 + c - w_i, \qquad (4)$$

зависящих нелинейно от искомых параметров,
где $c \equiv \frac{\sqrt{2\pi} \cdot \sigma \cdot \epsilon \cdot \alpha}{1-\epsilon}$.

Мы получили уравнения, первый член в которых аналогичен обычным уравнениям "взвешенного" МНК, но с заменой числовых весовых коэффициентов на *весовые функции*. Второй член полученных уравнений обусловлен принятой нами моделью загрязнения распределения ошибок измерений.

Таким образом, оптимальные веса w_i вычисляются через сложную экспоненциальную зависимость ошибок d_i, что, как хорошо известно, может приводить к неустойчивости при поиске экстремума функционала. Поэтому представляется целесообразным построить некоторые весовые функции, близкие к оптимальным, но менее резко убывающие с ростом ошибок и с более простой зависимостью от d_i. Как было отмечено в работе [12], полиномиальное разложение этих оптимальных весов до четвертого порядка приводит к биквадратной аппроксимации вида

$$w_{Tukey}(d_i) = \begin{cases} \left(1 - \frac{d_i^2}{\sigma^2 \cdot c_T^2}\right)^2 , d_i^2 \leq c_T^2 \cdot \sigma^2 \\ 0 \quad , d_i^2 > c_T^2 \cdot \sigma^2 \end{cases} , \qquad (5)$$

которая фактически является известными бивесами Тьюки [13], причем значения этих весов вычисляются значительно проще, чем значения оптимальных весов (параметр "обрезания" c_T обычно выбирают равным $3 \div 4$). На рис. 3 можно сравнить

[5] В работе [11] выполнена робастная оптимальная оценка параметров в условиях сильного загрязнения, однако для описания распределения ошибок измерений использовалась более простая суперпозиция распределений: нормального в центральной части и равномерного на концах.

поведение оптимальных весов и бивесов Тьюки в зависимости от относительных отклонений d_i/σ. Видно, что при такой аппроксимации удается достичь менее резкого убывания весов при больших отклонениях $|d_i|/\sigma > c_w = 2,54$ (как видно из рисунка, константа c_w вычисляется при условии $w_{optimal} = 0,4$). Однако можно заметить, что с ростом $|d_i|/\sigma$ от 0 до c_w бивеса Тьюки убывают более резко, чем оптимальные. Как выяснилось из проведенных расчетов, эта разница существенна, поскольку сказывается на результатах фитирования.

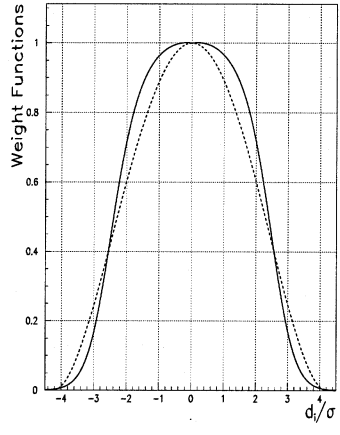

Рис. 3. Весовые функции в зависимости от относительных отклонений d_i/σ (оптимальные веса $w_{optimal}$ - сплошная линия; бивеса w_{Tukey} - пунктирная линия)

Поэтому таким же образом, как и для бивесов, нами была построена весовая функция 8-го порядка, которую мы именуем далее "квадро"весами (весовая функция 6-го порядка мало отличается от бивесов). Сравнительное поведение оптимальных и квадровесов показано на рис. 4. В отличие от бивесов, эти веса выглядят приемлемыми в области $|d_i|/\sigma < 2,54$, однако слишком резко убывают при больших отклонениях. Полученные результаты привели к естественному выводу использовать в качестве весовой функции бивеса при больших отклонениях и квадровеса — при малых [14]:

$$w_i = \begin{cases} [1 - (\frac{d_i}{c_4 \cdot \hat{\sigma}})^4]^2 , & |d_i| \le c_w \hat{\sigma} \ ; \\ [1 - (\frac{d_i}{c_2 \cdot \hat{\sigma}})^2]^2 , & c_w \hat{\sigma} < |d_i| \le c_2 \cdot \hat{\sigma} \ ; \\ 0 , & |d_i| > c_2 \cdot \hat{\sigma}, \end{cases} \qquad (6)$$

где $\hat{\sigma}^2 = \frac{\sum w_i \cdot d_i^2}{\sum w_i}$, $c_w = 2,54$, $c_4 = 3,26$ и $c_2 = 4,19$ [6].

[6]Константы c_2 и c_4 вычисляются из условия $w_{optimal} = w_{Tukey} = w_4 = 0,4$.

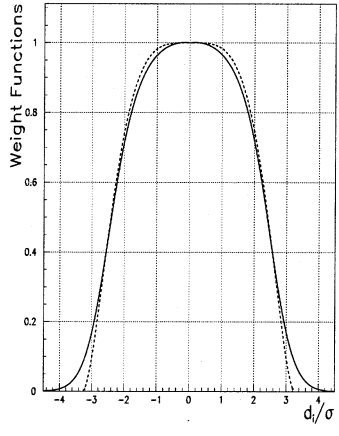

Рис. 4. Весовые функции в зависимости от относительных отклонений d_i/σ (оптимальные веса $w_{optimal}$ - сплошная линия; квадровеса w_4 - пунктирная линия)

Сравнение выбранных кусочно-непрерывных полиномиальных весов с оптимальными представлено на рис. 5.

Параметр $\hat{\sigma}$ получается из уравнения правдоподобия $\frac{\partial L}{\partial \sigma} = 0$. В итерационной форме это означает, что на k-й итерации параметр $\hat{\sigma}$ следует перевычислять, как рекомендовано в [12]:

$$\hat{\sigma}^{(k)^2} = \sum_i w_i^{(k-1)} (d_i^{(k-1)})^2 / \sum_i w_i^{(k-1)} . \tag{7}$$

Следуя Хьюберу [4], мы именуем этот подход порожденными М-оценками (descended M-estimates)[7]. Если нет никакой априорной информации, самое простое - начинать итерационный процесс с $w_i^{(0)} = 1$. Однако, как подчеркивается в [16], выбор начальных значений $w_i^{(0)}$ может играть немаловажную роль, и в данной работе предлагается следующая процедура выбора начальных значений весов:

- Перед нулевой итерацией мы применяем процедуру так называемой "базовой линии" для выбора первоначальных значений весов: составляя все возможные комбинации из p точек, проводим по ним базовую кривую (например, для прямой линии $p = 2$) с минимальной суммой модулей отклонений всех других

[7]В русском переводе с английского книги Хампеля и др. [10] эти оценки называются "сниженными". Справедливо критикуя этот перевод, Шурыгин [15] предложил назвать их "производными". Однако нам представляется, что поскольку этот термин уже имеет свой определенный математический смысл, то наиболее удачным русским эквивалентом "descended estimates" в данном контексте является выражение "порожденные оценки".

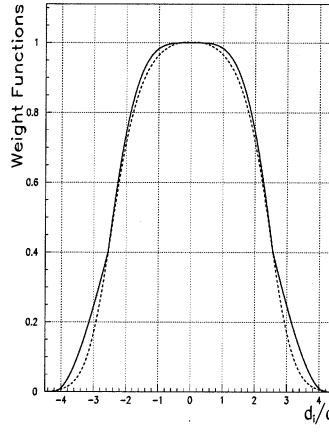

Рис. 5. Весовые функции в зависимости от относительных отклонений d_i/σ (кусочно-непрерывные полиномиальные веса (6) - сплошная линия; оптимальные веса $w_{optimal}$ - пунктирная линия)

измерений от этой кривой и задаем начальные веса для измерений, наиболее удаленных от "базовой линии", существенно меньше 1.

- После этого начальные веса умножаются на дополнительный весовой фактор, понижающий вес для измерений с малым значением заряда, индуцированного на стрипах, поскольку при наличии постоянного шума электроники точность измерения таких координат хуже по сравнению с измерениями, получаемыми по большим зарядам.[8]

Следует заметить, что в предлагаемом итерационном процессе также используется хорошо известная процедура "jack-knife" [13], которая называется также расщеплением выборки и состоит в проверке статистических выводов при отбрасывании поочередно двух (или более) точек с наибольшими, но близкими друг к другу отклонениями.

Таким образом, в данной работе приближенная итерационная процедура строится следующим образом: на нулевой итерации согласно процедуре "базовой линии" и с учетом величин зарядов на стрипах задаются начальные значения весов $w_i^{(0)}$, а значения $\tilde{w}_i^{(0)}$ - равными нулю. Тем самым система уравнений (3) приводится к линейной системе уравнений "взвешенного" МНК, из которой находятся значения искомых параметров $\theta_j^{(0)}$. Далее по известным параметрам вычисляются отклонения $d_i^{(0)}$, затем - $\hat{\sigma}^{(0)}$. Получив значения $d_i^{(0)}$ и $\hat{\sigma}^{(0)}$, приходим к

[8]См. ниже пояснение к рис. 9.

вычислению весов $w_i^{(1)}$ уже согласно формуле (6), а затем, соответственно, из (4) вычисляем $\tilde{w}_i^{(1)}$. Подставляя вычисленные веса в (3), получаем, что система уравнений (3) опять становится линейной по искомым параметрам $\theta_j^{(1)}$, и т.д. Выход из итерационного цикла осуществляется по пороговому значению для $\hat{\sigma}$. Практические вычисления показывают, что скорость сходимости - 2-3 итерации, что можно объяснить, в частности, удачно выбираемыми начальными условиями на нулевой итерации. Схематически итерационный процесс представлен на рис. 6.

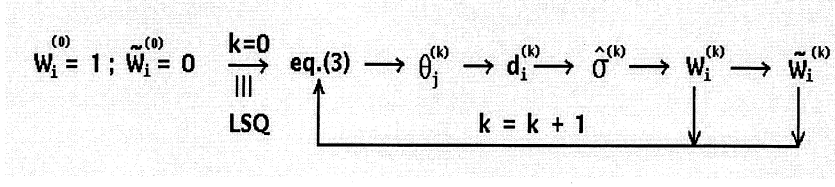

Рис. 6. Схема итерационной процедуры робастного метода фитирования

3 Модель Монте-Карло и полученные результаты

В работе построена математическая модель Монте-Карло (М.К.) линейной регрессии $x = az + b$ для прямолинейного (число параметров $p = 2$) мюонного трека, проходящего через 6 эквидистантных слоев CSC. Многократное рассеяние при этом не учитывается, поскольку в камерах мало вещества. Модель строится для прямолинейных треков, т. к. в работе используются экспериментальные данные с тестовых сеансов дубненского полномасштабного прототипа CSC, в которых мюонные траектории были прямолинейны. В условиях действующего эксперимента CMS траектории мюонных треков будут искривлены, однако распространить предлагаемый подход на параболическую траекторию не представляет трудностей, поскольку метод выведен для полиномов (1). Величина заряда, индуцированного на каждой из катодных плоскостей, моделировалась в соответствии с распределением Ландау [17]. Пространственное распределение заряда на катодной плоскости описывается формулой Гатти [18]. Суммарный заряд на каждом стрипе $q_j^{(0)}$ "размывался" (Δq_j) за счет добавления нормально-распределенного шума считывающей электроники с заданным стандартным отклонением σ_{noise} (т.е. $q_j^{(0)} \rightarrow q_j = q_j^{(0)} + \Delta q_j$). Поэтому восстановленный центроид распределения заряда x_i вычисляется с некоторой ошибкой.

При моделировании загрязнения также учитывалось вторичное электромагнитное сопровождение (δ-электроны), распределенное стохастически вдоль мюонного трека. Параметр загрязнения ϵ, число δ-электронов в каждом слое и расстояние между мюоном (x_i) и δ-электроном (x_e) (переменная экспоненциального распределения $|x_i - x_e|$) параметризуются на основе предшествующего моделирования физических процессов, происходящих при прохождении мюонов через вещество

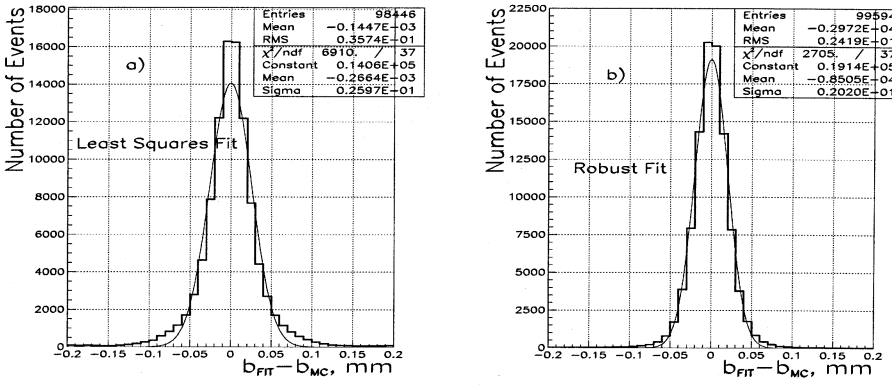

Рис. 7. Распределения отклонений оценок параметров сдвига прямой линии b_{FIT} от М.К.-параметров b_{MC}: a) МНК; b) робастный подход. Распределения $b_{FIT} - b_{MC}$ профитированы гауссианом, что изображено на рисунках тонкими линиями

калориметра и CSC, с применением программы GEANT [9]. Для сравнения было также проделано упрощенное моделирование загрязнения с использованием широкого гауссиана для распределения $(x_i - x_e)$ с $\sigma \approx 1$мм, как это описано в [19].

Как видно из рис. 7, распределение отклонений свободного члена регрессии (параметра сдвига) b для фитирования по МНК[9] имеет более длинные "хвосты", чем распределение, полученное в робастном подходе. Среднеквадратичное отклонение (RMS) существенно изменяется при наличии выбросных точек[10]. В общепринятой практике физики рассматривают RMS как основной показатель качества фитирования. Поэтому в случае использования МНК при обработке данных наиболее удаленные точки игнорируются (вычеркиваются), что приводит к уменьшению значения RMS. Однако в робастном подходе такого отбрасывания точек не требуется. Уменьшение влияния выбросов происходит автоматически с помощью робастных весов (6). Конечно, в этой формуле имеется параметр обрезания c_2, что ведет к потере около 0,4 % событий (см. "Entries" на рис. 7a и 7b), содержащих искаженные загрязнением координаты в трех и более плоскостях. Однако это производится более "щадящим" образом, чем прямое отбрасывание точек, часто используемое при фитировании по МНК, и, как видно из рис. 7a, потери ($\approx 1,6\%$) в 4 раза больше.

Отбросив при фитировании по МНК в 4 раза больше событий, чем в робастном подходе, казалось бы, можно рассчитывать значительно уменьшить RMS. Тем не менее параметры треков, полученных робастным подходом (рис. 7b), имеют

[9]Как сказано выше, в работе используется МНК с вычеркиванием выбросов и перефитированием, как описано в [5].

[10]Для того чтобы продемонстрировать влияние на RMS не просто отдельных выбросных точек, а получить статистически значимые отклонения, М.К. расчеты были проведены на достаточно больших выборках (около 100 тысяч событий в каждой выборке).

величину RMS в 1,5 раза лучше, чем параметры, полученные при фитировании по МНК. Для параметра наклона a получаются аналогичные результаты.

Как это хорошо известно из математической статистики (см., например, [20, 21, 24]), любая оценка параметров может быть квалифицирована доверительным уровнем. В нашем случае процент событий, в которых хотя бы один из параметров (a, b) находится вне 95 %-го совместного доверительного интервала, составляет 4,9% для робастного фитирования трека и 22,7 % для фитирования по МНК. Число 22,7 % настолько статистически значимо, что явно свидетельствует о непригодности МНК для фитирования треков по загрязненным экспериментальным данным.

Если варьировать уровень шума электроники в М.К.-модели, то можно получить величину среднеквадратичных отклонений для параметра сдвига RMS($b_{fit} - b_{MC}$) в зависимости от шума. Результаты вычислений как для загрязненных, так и для незагрязненных данных представлены на рис. 8, откуда видно, что соотношение RMS(robust)/RMS(LSQ) примерно одинаково на данных без электромагнитного сопровождения, а на данных с загрязнением составляет приблизительно 1,4 при малом шуме (1 ADC), 1,6 - при высоком уровне шума, причем в рабочей области (2-6 ADC) это соотношение достигает величины 1,7.

Здесь необходимо небольшое пояснение. Как можно видеть из рис. 8, робастное фитирование дает несколько лучший результат по сравнению с МНК даже в случае незагрязненных данных. То есть различие между RMS_{LSQ} и RMS_{robust} возрастает до 5 мкм для параметра сдвига при $\sigma_{noise} = 12$ ADC. На первый взгляд может показаться, что имеет место нарушение фундаментальной теоремы Гаусса–Маркова (см., например, [22]), согласно которой на незагрязненных данных МНК должен давать наилучшую оценку параметров с минимальной дисперсией. Однако если рассмотреть данную ситуацию детально, то следует заметить, что точность реконструкции центроида индуцированного заряда зависит от соотношения сигнал/шум, которое не является постоянным. На рис. 9a для $\sigma_{noise} = 4$ ADC показана зависимость разрешения от индуцированного заряда для различных интервалов по заряду. Как можно видеть, в каждом малом интервале зарядов величины RMS и σ очень близки друг к другу, что говорит о том, что распределение ошибок очень близко к гауссиану[11]. На рис. 9b показано общее распределение разницы между измеренной и идеальной координатами $(x_c - x_{MC})$ для всех зарядов. Это распределение фактически является распределением ошибок измерений. Если профитировать его гауссианом, можно видеть существенное различие между гауссианом и суммарным распределением ошибок. Такой результат объясняет упомянутое выше расхождение с теоремой Гаусса–Маркова. Однако, если зафиксировать заряд (возьмем, к примеру, $Q_{fixed} = 480$ ADC) при моделировании данных без загрязнения, мы получим (как и ожидаемо для гауссова распределения ошибок), что МНК дает наилучшую оценку параметров в полном соответствии с теоремой Гаусса–Маркова. При этом RMS_{LSQ} оказался на 3 мкм лучше, чем в робастном подходе.

Испытывая различные модели загрязнения, легко проверить гипотезу [23] о вы-

[11]Увеличение разницы между RMS и σ с увеличением индуцированного заряда Q_{clust} объясняется просто уменьшением статистики в соответствии с распределением Ландау, показанным на рис. 9a в виде гистограммы.

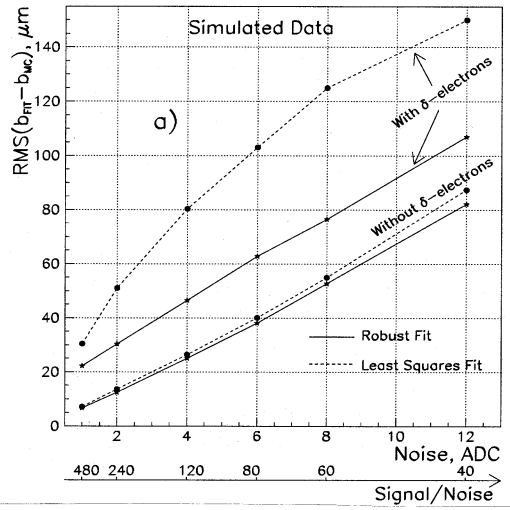

Рис. 8. Зависимость среднеквадратичных отклонений RMS ($b_{fit} - b_{MC}$) для параметра сдвига от различных уровней шума электроники (рабочая область находится между 2 и 6 ADC) для МНК и робастного фитирования, полученная на модельных данных с учетом и без учета загрязнения: пунктирные линии - фитирование по МНК; сплошные линии - робастное фитирование

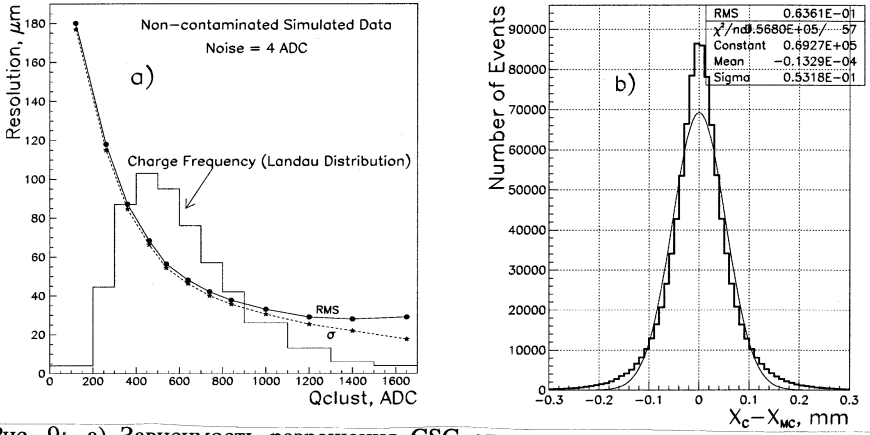

Рис. 9: a) Зависимость разрешения CSC от величины индуцированного заряда Q_{clust} (просуммированного по стрипам) для $\sigma_{noise} = 4$ ADC на незагрязненных модельных данных. RMS разностей между восстановленной ("измеренной") координатой x_c и идеальной координатой x_{MC} показана сплошной линией; σ этого распределения, профитированная гауссианом по каждому зарядовому интервалу, показана пунктирной линией. Гистограмма иллюстрирует частоту появления кластеров в том или ином зарядном интервале в соответствии с распределением Ландау; b) Общее ($x_c - x_{MC}$)-распределение для всех зарядов, дающее усредненные значения RMS и σ

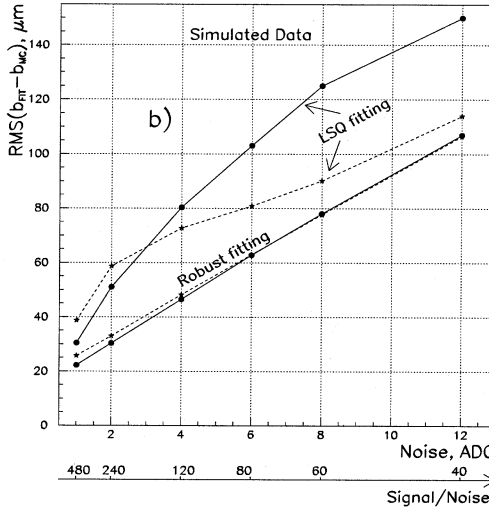

Рис. 10. Зависимость RMS ($b_{fit} - b_{MC}$) от шума электроники на модельных данных для различных моделей загрязнения: пунктирные линии - для модели загрязнения в виде широкого гауссова распределения ($x_i - x_e$); сплошные линии - для двусторонней экспоненциальной модели загрязнения

сокой адаптивности робастных оценок к основному распределению ошибок. На рис. 10 представлены RMS($b_{fit} - b_{MC}$) для двух моделей загрязнения: двусторонней экспоненциальной и модели в виде широкого гауссиана. Можно видеть, что результаты робастного фитирования, полученные для первой модели, в основном совпадают для обеих моделей (они различаются не более чем на несколько процентов), а для результатов фитирования по МНК эта разница может достигать 30 %. Этот результат демонстрирует действительно высокую адаптивность робастного фитирования к различным типам загрязнения.

На практике часто необходимо определить разрешение прибора с использованием результатов фитирования. Для оценки пространственного разрешения следует вычислить стандартизированные остатки [24] на модельных данных. Применяя МНК, мы получаем распределение остатков, которое отличается от гауссиана и имеет длинные "хвосты" (см. рис. 11а). Во-первых, можно заключить, что RMS-распределения существенно выше (приблизительно в 1,3 раза), чем модельное разрешение прибора, которое составляет ≈ 65 мкм при уровне шума электроники 4 ADC. Затем можно сделать попытку оценить гауссову компоненту (которая соответствует пространственному разрешению камеры), фитируя это распределение

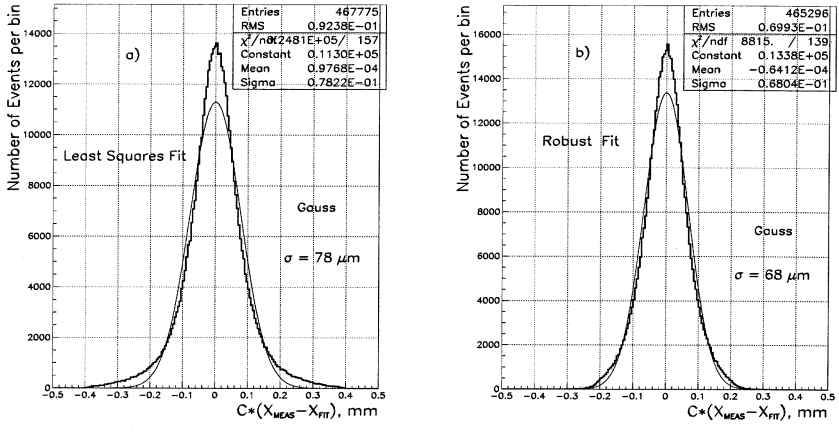

Рис. 11. Распределения стандартизированных остатков для МНК и робастного фитирования треков на загрязненных модельных данных с шумом электроники $\sigma_{noise} = 4ADC$ (профитировано гауссианом)

различными способами, что приводит нас к различным результатам (см. рис. 12) и, соответственно, к широкой неопределенности в оценке разрешения. Напротив, распределение остатков, полученное при робастном фитировании (рис. 11b), очень близко к гауссиану ($RMS \approx \sigma$), и, таким образом, можно сделать однозначный вывод о пространственном разрешении камеры, близкий к модельному.

После проверки предлагаемого робастного подхода для реалистичной оценки разрешения прибора на модельных данных мы применили этот метод к экспериментальным данным, полученным на прототипе детектора [5]. Эти данные были получены на дубненском полномасштабном прототипе CSC на установке Интегральный Тест, помещенной на пучке H2 на ускорителе SPS в ЦЕРН. Съем данных производился с мюонного пучка с импульсами от 100 до 300 ГэВ/с. Поскольку CSC была расположена за адронным калориметром, мюонные измерения были сильно загрязнены δ-электронами и вторичным электромагнитным сопровождением.

Сравнительный пример результатов фитирования прямолинейного мюонного трека на экспериментальных данных с дубненского прототипа CSC методом наименьших квадратов и робастным методом изображен на рис. 13. Как хорошо видно из рисунка, далеко удаленная точка в 6-й плоскости вычеркнута из процедуры фитирования в обоих методах. В то же время выброс в 1-й плоскости (outlier) при минимизации функционала квадратичных отклонений по МНК (LSQ) приводит к притягиванию линии регрессии к этому измерению. При этом среднеквадратичное отклонение экспериментальных точек от регрессионной прямой (RMS_{LSQ}) оказалось меньше порогового значения ($RMS_{THRESHOLD}$), и процедура фитирования по МНК на этом заканчивается. Однако заметим, что если бы порог был выбран меньше, то наиболее удаленной точкой, подлежащей вычерки-

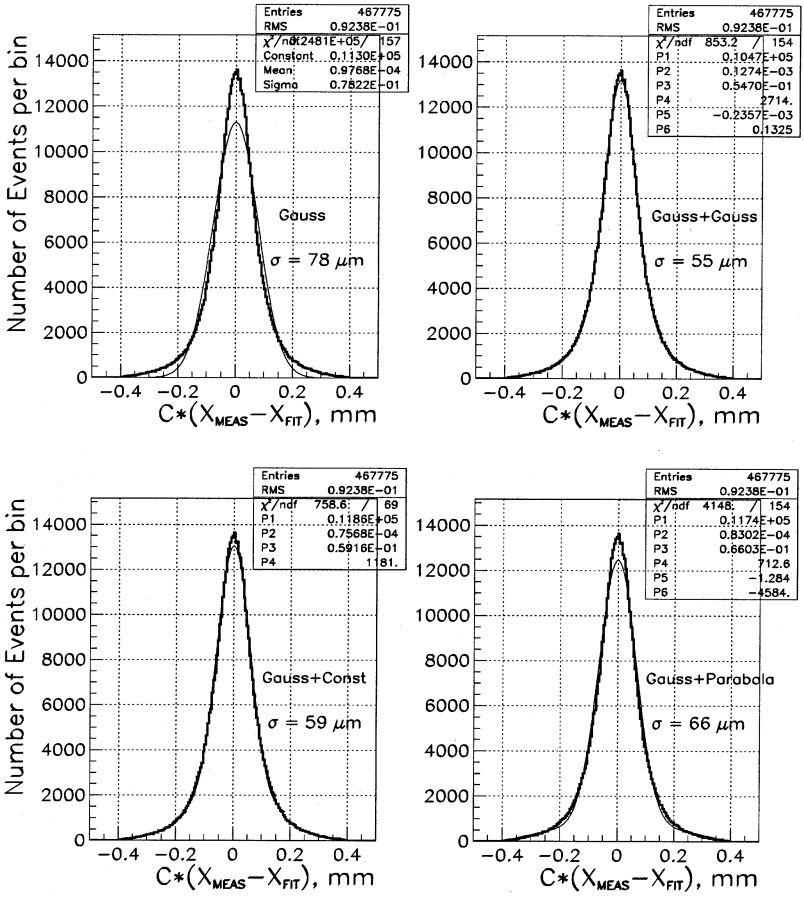

Рис. 12. Распределения стандартизированных остатков для фитирования треков по МНК на модельных данных (4 варианта фитирования распределений остатков: гауссианом, двумя гауссианами, гауссианом и константой, гауссианом и параболой)

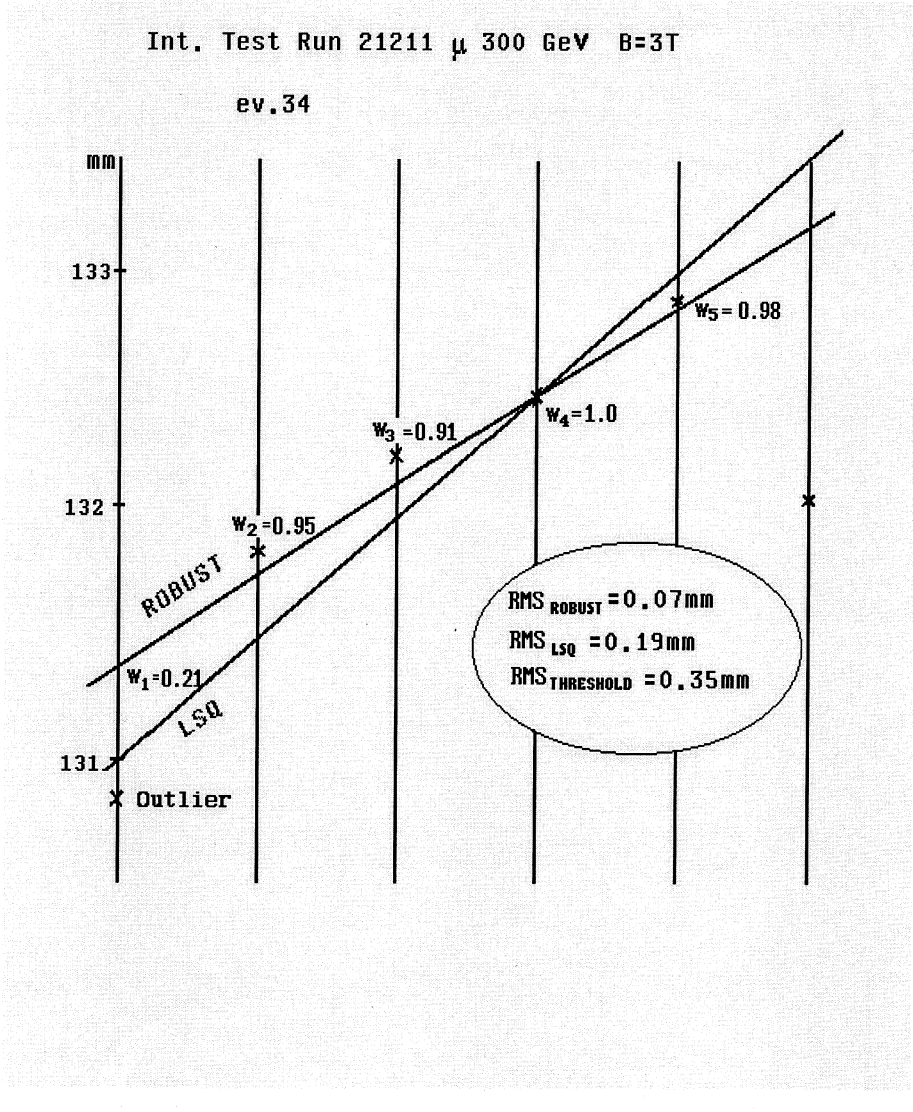

Рис. 13. Пример результата фитирования траектории мюона на экспериментальных данных с прототипа CSC

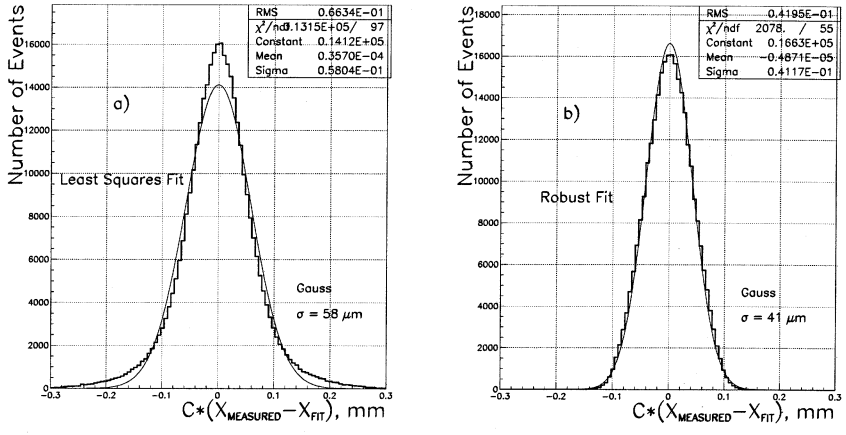

Рис. 14. Распределения стандартизированных остатков для МНК и робастного фитирования треков на экспериментальных данных (профитировано гауссианом)

ванию в МНК, оказалось бы измерение во 2-й плоскости, которое, как хорошо видно даже "на глаз", следовало бы сохранить. Совершенно иначе происходит фитирование в робастном подходе (ROBUST), где уменьшением веса ($w_1 = 0,21$) выбросной точки ослабляется ее влияние и удается провести регрессионную прямую в непосредственной близости от измерений во 2, 3, 4 и 5-й плоскостях, причем среднеквадратичное отклонение (RMS_{ROBUST}) получилось в 2,7 раза лучше, чем при фитировании по МНК.

Результаты, полученные на экспериментальных данных с применением МНК и робастного фитирования, показаны на рис. 14. Нетрудно увидеть значительные негауссовы "хвосты" в распределении остатков для фитирования по МНК, в то время как распределение остатков для робастного фитирования достаточно близко к гауссиану. Таким образом, можно сделать вывод, что предлагаемый робастный подход может быть использован для более достоверной оценки пространственного разрешения CSC, которое для данного прототипа камеры в средней ее части (области мюонного пучка) составляет 41 мкм.

4 Заключение

Для обработки данных с мюонной камеры при условиях сильного загрязнения вместо общепринятого метода наименьших квадратов предложен метод робастного фитирования треков, который по сути является итерационным взвешенным методом наименьших квадратов с кусочно-непрерывной полиномиальной аппроксимацией оптимальных весовых функций, уменьшающих влияние выбросов на результаты. В работе оптимальные весовые функции выведены аналитически с

учетом реалистического загрязнения данных вследствие электромагнитного сопровождения мюонов.

Для проверки предлагаемой процедуры фитирования треков создана программа моделирования данных в CSC с учетом загрязнения данных в соответствии с условиями физического эксперимента.

Расчеты на модельных данных показывают, что параметры треков, полученные с применением робастного подхода, имеют среднеквадратичное отклонение в 1,5-1,7 раза лучше, чем параметры, полученные с помощью МНК.

Более того, следует отметить, что, в то время как 5 % событий, полученных по робастному методу, лежат вне 95 %-го доверительного интервала по параметрам, для фитирования по МНК доля таких событий превышает 20 %. Таким образом, мы приходим к выводу, что для данных с реалистическим загрязнением статистическая эффективность робастного подхода остается высокой, в то время как в МНК на этих данных получено значимое отклонение распределений по параметрам треков от распределения Стьюдента, и тем самым возможность использования МНК на загрязненных данных должна быть отвергнута.

Длинные негауссовы "хвосты" в распределениях МНК-остатков для CSC-данных с большим загрязнением ведут к широкой неопределенности в оценке пространственного разрешения CSC. Распределения робастных остатков, напротив, очень близки к гауссиану.

Исследованием различных моделей загрязнения данных в работе убедительно продемонстрирована высокая адаптивность робастного фитирования к различным типам загрязнения. В то же время проделанные вычисления показывают высокую чувствительность фитирования по МНК к различным типам загрязнения.

Полученные результаты определенно доказывают необходимость использования робастной процедуры фитирования для оптимальной оценки параметров трек-сегментов в катодно-стриповых камерах CMS, а также для достоверного вычисления пространственного разрешения мюонных камер на реальных данных.

Список литературы

[1] CMS Collaboration. The Compact Muon Solenoid, Technical Proposal, CERN/LHCC 94-38, LHCC/P1, CERN, 1994.

[2] С.А.Айвазян, И.С.Енюков, Л.Д.Мешалкин. Прикладная статистика. Исследование зависимостей, М.: Финансы и статистика, 1985.

[3] С.А.Смоляк, Б.П.Титаренко. Устойчивые методы оценивания (Статистическая обработка неоднородных совокупностей), М.: Статистика, 1980.

[4] P.Huber. Robust Statistics, J.Willey&Sons, NY, 1981; П.Хьюбер. Робастность в статистике, М.: Мир, 1984.

[5] I.A.Golutvin et al. Muon track reconstruction efficiency of ME1/1 prototype in the Integrated Test, CMS Note/ **1997-084** , CERN, Geneva, 1997.

[6] I.A.Golutvin, Y.T.Kiriouchine, S.A.Movchan, G.A.Ososkov, V.V.Palichik, E.A.Tikhonenko. Robust estimates of track parameters and spatial resolution for CMS muon chambers//Computer Physics Communications, **126(2000)**, 2000, pp. 72-76.

[7] Вероятность и математическая статистика. Энциклопедия. Гл. ред. Ю.В.Прохоров, М.: Большая Российская энциклопедия, 1999.

[8] J.W.Tukey. A survey of sampling from contaminated distribution//Contribution to Probability and Statistics. Ed. I.Olkin, Stanford: Stanford Univ. Press, 1960, pp.446-486.

[9] GEANT - Detector Description and Simulation Tool. CERN Program Library Long Writeup W5013, CERN, Geneva, 1993.

[10] Ф.Хампель, Э.Рончетти, П.Рауссеу, В.Штаэль. *Робастность в статистике*, М.: Мир, 1989.

[11] N.I.Chernov, G.A.Ososkov. Joint estimates of location and scale parameters, JINR Preprint **E10-86-282** , Dubna, 1986.

[12] G.Agakishiev et al. Cherenkov ring fitting techniques for the CERES RICH detectors//Nuclear Instruments and Methods **A 371**, 1966, p.243.

[13] F.Mosteller, J.W.Tukey. *Data analysis and regression: a second course in statistics*, Addison - Wesley, NY, 1977; Ф.Мостеллер, Дж.Тьюки. Анализ данных и регрессия, М.: Финансы и статистика, 1982.

[14] I.Golutvin, S.Movtchan, G.Ososkov, V.Palichik, E.Tikhonenko. Optimal Choice of Track Fitting Procedure for Contaminated Data in High-Accuracy Cathode Strip Chambers// Proc. of CHEP'2000, Padova, Italy, 2000, pp. 128-131.

[15] А.М.Шурыгин. Прикладная стохастика: робастность, оценивание, прогноз, М.: Финансы и статистика, 2000.

[16] У.С.Аджи, Р.Х.Тернер. Применение методов помехоустойчивого оценивания в анализе данных о траекториях движения//Устойчивые статистические методы оценки данных/ под ред. Р.Л.Лонера и Г.Н.Уилкинсона. М.: Машиностроение, 1984, с.86-105.

[17] Л.Д.Ландау. О потерях энергии быстрыми частицами на ионизацию. J.Phys.USSR, 1944, vol.8, p.201; Собрание трудов, т.1. М.: Наука, 1969. с.482.

[18] E.Gatti et al. Optimum Geometry for Strip Cathodes or Grids in MWPC for Avalanche Localization along the Anode Wires. Nuclear Instuments and Methods// **163**, 1979, pp.83-92.

[19] Y.Kiriouchine, S.Movchan, G.Ososkov, V.Palichik, E.Tikhonenko. Estimation of CSC Prototype Spatial Resolution by Robust Method// Proc. of Third RMDS CMS Annual Meeting, **1997-168**, CERN, Geneva, 1998, pp.363-370.

[20] H.Cramer. Mathematical Methods of Statistics, Stockholm, 1946; Г.Крамер. Математические методы в статистике, М.: Мир, 1975.

[21] D.E.Groom et al. Review of Particle Physics, Eur. Phys. J. **C 15**, 2000, p.198.

[22] D.J.Hudson. Statistics Lectures II, **64-18**, CERN, Geneva, 1964; Д.Худсон. Статистика для физиков, М.: Мир, 1970.

[23] R.V.Hogg. Adaptive Robust Procedure: A Partial Review and Some Suggestions for Future Applications and Theory//Journal of the American Statistical Association. Vol.69, No.348, 1974, pp.909-927.

[24] М.Кендал, А.Стюарт. Статистические выводы и связи, М.: Наука, 1973.

Рукопись поступила в издательский отдел
18 июля 2001 года.

Голутвин И.А. и др. P13-2001-147
Робастные оптимальные оценки параметров трек-сегментов мюонов в катодно-стриповых камерах эксперимента CMS

Катодно-стриповые камеры (CSC) будут использоваться в мюонной системе строящейся в ЦЕРН установки CMS. CSC должны обеспечивать высокую точность (~ 100 мкм) измерения пространственных координат μ-мезонов в тяжелых фоновых условиях, когда ~ 20 % измеренных координат мюонов зашумлено вторичным электромагнитным сопровождением и во многих случаях перекрывающиеся срабатывания разделить невозможно. Вследствие такого загрязнения данных распределение ошибок измерений существенно отличается от нормального (гауссова). В подобных случаях использование традиционного метода наименьших квадратов становится необоснованным.

В работе предлагается робастная (помехоустойчивая) итерационная процедура фитирования трек-сегментов в CSC. Выведена в аналитическом виде оптимальная весовая функция с учетом реалистического загрязнения за счет электромагнитного сопровождения. Для ускорения итерационной процедуры получена удачная кусочно-непрерывная аппроксимация оптимальной весовой функции. Проведен сравнительный анализ расчетов как на модельных, так и на экспериментальных данных с прототипа камеры. Полученные результаты определенно доказывают необходимость использования робастного подхода при фитировании треков в CSC на данных с сильным зашумлением.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.
Препринт Объединенного института ядерных исследований. Дубна, 2001

Golutvin I.A. et al. P13-2001-147
Robust Optimal Estimates of Muon Track Segment Parameters in Cathode Strip Chambers of the CMS Experiment

Cathode strip chambers (CSC) will be used as muon detectors in a forward region of the compact muon solenoid (CMS) setup which is being constructed at CERN. CSC should provide a high accuracy (~ 100 μm) of measurements of muon space coordinates under conditions of heavy background when ~ 20 % of measured coordinates are contaminated with a secondary electromagnetic accompaniment and in many cases double hits cannot be separated. Due to these data contaminations the error distribution of the muon coordinate measurements differs from a normal distribution. In such a case the usage of the conventional least squares method (LSQ) for data processing becomes groundless.

A robust iterative procedure of track fitting in CSC has been proposed. An optimal weight function has been deducted analytically taking into account a realistic contamination due to an electromagnetic accompaniment. A proper piecewise continious approximation has been obtained in order to speed up the iterative procedure. A comparative analysis has been done on both simulated data and experimental measurements from the chamber prototype. The results obtained show definitely a necessity of the usage of a robust method for track fitting in CSC under conditions of heavy background.

The investigation has been performed at the Laboratory of Information Technologies, JINR.
Preprint of the Joint Institute for Nuclear Research. Dubna, 2001

Редактор А.Н.Шабашова. Макет Р.Д.Фоминой

Подписано в печать 18.09.2001
Формат 60 × 90/16. Офсетная печать. Уч.-изд. л. 1,95
Тираж 330. Заказ 52860. Цена 2 р. 34 к.

Издательский отдел Объединенного института ядерных исследований
Дубна Московской области