Метод эластичных нейронных сетей и его робастная трактовка

Ососков Г.А.

Объединенный институт ядерных исследований, Дубна, Московской обл., soskov@jinr.ru

Аннотация. После краткого обзора методов обработки данных современных экспериментов физики высоких энергий описаны методы эластичных нейронных сетей, которые затем трактуются как частный случай робастного (устойчивого к засорениям выборки) подхода к оцениванию зависимостей. Описан формализм робастного подхода с выводом оптимальных весов для случаев равномерного засорения выборки. На базе исследования совместной оценки параметров положения и масштаба получен экономичный алгоритм робастного оценивания.

1. Введение

Распознавание кривых линий на сложном фоне шумовых точек и близких соседних кривых является классической проблемой распознавания образов во многих важных научно-технических приложениях. Особенно сложными такие проблемы оказываются в ядерной физике высоких энергий, где по данным измерений необходимо не только распознать траекторию отдельной заряженной частицы среди тысяч траекторий других частиц, но и дать оценку ее параметров, в том числе с максимально возможной точностью вычислить импульс частицы, определяемый кривизной ее траектории в магнитном поле. Хотя эти проблемы распознавания и последующей подгонки траекторий (треков) заряженных частиц имеют почти полувековую историю, методы их решения существенно менялись по мере эволюции экспериментальных установок от пузырьковых, где несколько треков регистрировались на стереофотографиях, до современных экспериментов с тяжелыми ионами, в которых рождаются тысячи треков, фиксируемых быстрыми электронными детекторами прямо в памяти компьютеров в виде массивов измеренных координат (см. рис.1). К тому же кругу проблем относится и обработка данных с детекторов черенковского излучения типа RICH, в которых черенковские фотоны, образующиеся при пролете частицы со скоростью выше скорости света в данной среде, отражаются от сферического зеркала на фотоматрицу, где они регистрируются в виде характерных колец (см. рис. 2).

Главные требования к обработке в современных экспериментах: максимальная скорость вычислений при предельно достижимой их точности и высокая эффективность методов оценки физических параметров, интересующих экспериментаторов. Реализация этих требований при наличии чрезвычайно сложной текстуры и зашумленности экспериментальных данных неизбежно натолкнулась на ограниченность традиционно применяемых классических комбинаторных методов, кластерного анализа и подгонки по методу наименьших квадратов, которые в этих условиях уже не обеспечивали либо точности, либо

скорости вычислений, либо высокой эффективности оценок параметров, либо всего этого вместе [1].

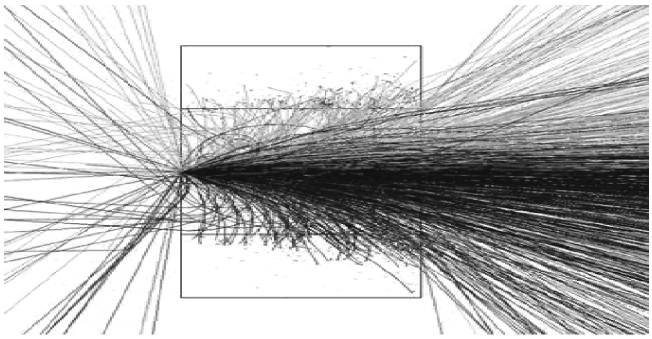

Рис.1. Смоделированное событие множественного рождения при центральном столкновении атомов золота при энергии 25 AGeV/c. Эксперимент CBM, Германия.

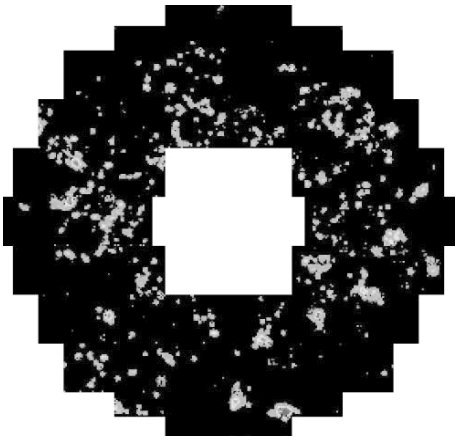

Рис.2. Снимок с фотоматрицы RICH детектора с изображениями колец черенковского излучения, зарегистрированных на шумовом фоне. Эксперимент CERES/NA-45, CERN.

Возьмем в качестве примера проблему распознавания треков. Если раньше при множественности детектируемых событий в несколько треков на событие основной методикой распознавания было «прослеживание в дороге» (Road Guidance – RG, см. статью Р.Стренда [2]), т.е. кластеризация соседних точек по угловой близости отрезков, их соединяющих, то для множественности в несколько сот треков на событие потребовались новые, более совершенные методы распознавания. Вполне естественным было обращение к аппарату искусственных нейронных сетей (ИНС). Однако помимо многих нейросетевых (и поэтому быстрых) реализаций того же RG-метода с помощью многослойных персептронов [3], принципиально новым явился предложенный Денби и Петерсоном (ДП) метод, основанный на использовании полносвязных нейронных сетей

типа Хопфилда [4]. В ДП методе трек рассматривается как последовательность сегментов, соединяющих соседние измеренные точки. Идея метода состоит во введении бинарных нейронов S_{ij}, равных единице, если трек соединяет точки i и j , и нулю в противном случае. Принадлежность точек треку, как к некой гладкой кривой, обеспечивалась убывающей функциональной зависимостью синаптических весов w_{ijkl} для пары нейронов S_{ij} и S_{kl} от угла между сегментами трека, соответствующими этим нейронам, и от расстояния между этими сегментами. Энергетическая функция ИНС как билинейный функционал от S_{ij} и S_{kl} строилась так, чтобы набор гладких треков соответствовал ее минимуму. Несмотря на очевидную чувствительность такого метода к наличию шумовых измерений, ДП метод и аналогичный подход с так называемыми роторными ИНС [5], благодаря применению различных обобщений и системы пороговых обрезаний нашли весьма широкие применения в экспериментальной физике [6,7].

Тем не менее, как было продемонстрировано в исследовании [8], в современных экспериментах с множественностью в несколько тысяч треков на событие нейронные сети, как непосредственный инструмент распознавания треков, перестают удовлетворять экспериментаторов. Это может быть объяснено прежде всего тем, что при использовании методов типа ДП не применяется информация о параметрической форме трека, которая обычно известна заранее. ИНС используется только для решения комбинаторной задачи поиска точек и присоединения их к «своему» треку, в то время как, например, учет его локальной кривизны, известной из его уравнения, мог бы позволить сделать этот поиск более надежным. Недаром столь удачным в приложениях оказался опыт учета параметрической формы трека в синаптических весах роторной ИНС, позволивший выполнить обработку реальных экспериментальных данных [9] и провести анализ ионограмм [10].

Идея объединения этапов прослеживания и фитирования треков была предложена в работе [11] в виде **эластичной нейронной сети**, использующей гибкий шаблон, т.е. уравнение трека, зависящего от вариаций параметров таким образом, чтобы, изгибаясь при их изменении, кривая, описываемая этим уравнением (ее еще называют: *деформируемый шаблон, эластичная рука* и т.д.), прошла как можно ближе к «своим» точкам, измеренным на треке, соответствующем этому шаблону, «отталкиваясь» от «чужих». В качестве средства для реализации этой идеи были взяты бинарные нейроны, равные единице, если точка принадлежала шаблону и нулю в противном случае. Минимизация энергетической функции сети выполнялась при ограничении, учитывающем наличие шумовых точек, не принадлежащих никакому шаблону,

В данной статье после изложения метода эластичных нейронных сетей и близкого к нему подхода [8], также реализующего идею о совмещении этапов поиска и фитирования кривых, дано краткое введение в теорию робастных методов подгонки и показано, что оба метода «эластичного» трекинга могут рассматриваться как частные случаи робастного подхода.

2. Эластичные нейронные сети

Как отмечалось выше, идея работы [11] по объединению этапов распознавания и фитирования кривых привела к подходу учета известного уравнения трека путем создания эластичного шаблона и использования нейронов для сортировки точек. Такой подход немедленно порождал естественные вопросы о неизвестном число шаблонов, инициализации их параметров и об организации подгонки сразу всех кривых (треков).

Проблема нахождения шаблонов (templates) и грубых начальных значений их параметров требует специального рассмотрения. Мы ограничимся здесь ссылкой на работы [11-13], в которых использовались эластичные шаблоны, и отметим, что всюду для их поиска применяется тот или иной вариант преобразования Радона-Хафа [14,15]. При практической реализации это преобразование сводится к гистограммированию в пространстве параметров с последующим поиском максимального значения. В силу приближенного характера процедуры нахождения шаблонов их число могло превышать число реальных треков в событии. Появляющиеся из-за этого лишние, искусственные треки должны быть удалены впоследствии в ходе специальной процедуры отбраковки.

Для поиска сразу всех треков, порожденных взаимодействием в однородном магнитном поле, в работе [11] предложено минимизировать следующий функционал:

$$E(\{S_{ia}\};\vec{\pi}) = \sum_{i=1}^{N}\sum_{a=1}^{M} S_{ia} D_{ia}(\vec{x},\vec{\pi}) + \lambda \sum_{i=1}^{N}\left(\sum_{a=1}^{M} S_{ia} - 1\right)^2 . \qquad (1)$$

Здесь $\boldsymbol{\pi}$ – вектор параметров винтовой линии (геликоиды), описывающей траекторию движения a-й частицы в пространстве; S_{ia} – бинарный нейрон, определяющий принадлежность i-й точки к a-му шаблону трека по правилу: $S_{ia} = 1$, если i-я точка принадлежит a-му треку, $S_{ia} = 0$, в противном случае, $D_{ia}(\mathbf{x}, \boldsymbol{\pi})$ - квадрат расстояния от точки $\mathbf{x}$ до a-го трека.

Минимизация $E(S_{ia};\boldsymbol{\pi})$ ведется при условии, что каждая точка может принадлежать только одному треку или ни одному ($\Sigma_{ia} S_{ia} = 0$). В последнем случае функционал штрафуется на величину λ, что определяет критическое расстояние $\sqrt{\lambda}$, до которого точки энергетически выгодно включать в трек, а после – считать такую точку шумовой ($S_{ia} = 0, \ \forall\, i$).

Дальнейшие вычисления по поиску глобального минимума функционала (1) ведутся в соответствии с обычной схемой сетей Хопфилда: сеть термализуется (используется метод отжига [17], т.е. вводится модельная температура T, характеризующая случайный шум в процессе поиска минимума функционала (1), и T постепенно уменьшается) и применяется теория среднего поля. Для последовательности уменьшающихся температур $T_k > T_{k\text{-}1} >...> T_0$ метод наискорейшего спуска дает пошаговый итерационный алгоритм поиска минимума [11]:

$$\Delta\pi_a^{(k)} = -\eta_a^{(k)} \sum_i V_{ia} \frac{\partial D_{ia}}{\partial \pi_a^{(k)}} \ , \tag{2}$$

где

$$V_{ia} = \frac{e^{-\beta D_{ia}}}{e^{-\beta \lambda} + \sum_{b=1}^{M} e^{-\beta D_{ib}}} \tag{3}$$

– факторы Поттса, $\beta = 1/T$.

Метод эластичных нейросетей был успешно применен в наших работах [12,13] для обработки данных, параметризуемых уравнением окружности: колец черенковского излучения и треков в однородном магнитном поле.

В первой работе [12] методом эластичных нейросетей (ЭН) осуществлялся поиск черенковских колец с одновременной оценкой их параметров по данным RICH детектора CERES/NA-45 (см. пример на рис.3). Глобальный поиск велся безо всяких априорных сведений о центрах и радиусах колец. Эта начальная информация была получена с помощью сведения трехмерного преобразования Хафа к двумерному и одномерному. Вначале выполнялся перебор допустимых триплетов измеренных точек (т.е. таких троек точек, через которые можно провести окружность с радиусом и центром, удовлетворяющим заданным неравенствам). Идея метода основана на том, что все точки, принадлежащие некой окружности, должны отображаться в одну точку в пространстве параметров, так что, суммируя, мы должны получить пик в этом месте. Примеры работы алгоритма в этих двух режимах представлены на рис.3. Отсев ложных колец, получившихся из-за того, что преобразование Хафа нашло несколько лишних шаблонов (см. пример в левом нижнем углу рис.3), мог быть легко выполнен в последующем анализе с использованием критерия χ^2 и учетом статистики по числу и угловому распределению фотонов в кольце.

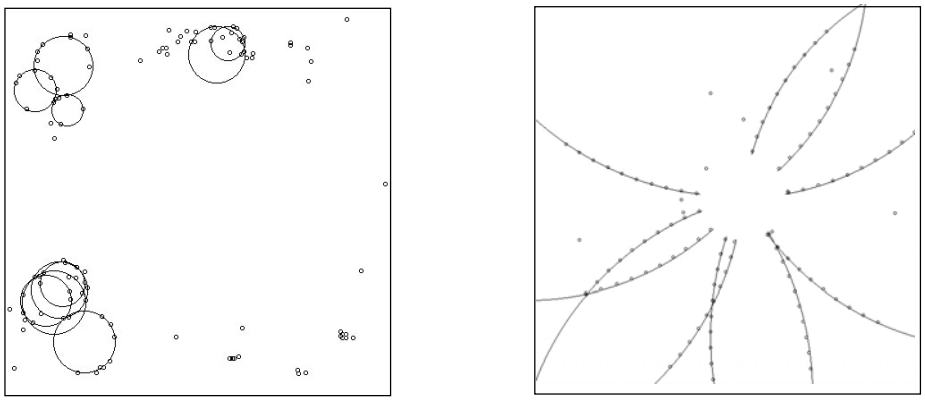

Рис.3. Примеры результатов работы метода эластичных нейросетей из [12]: слева – распознавание черенковских колец; справа – распознавание треков в магнитном поле.

Следует, однако, сделать важное замечание, касающееся издержек глобальности метода ЭН: при попытках увеличить число колец или множественности, т.е. числа треков, в районе поиска выше 12-13, метод переставал работать (не говоря уж о быстром росте машинного времени). Частично это происходило из-за ошибок в преобразовании Хафа, но главным образом из-за проблем с минимизацией функционала.

Поэтому во второй работе [13], касавшейся применения метода ЭН в задаче распознавания и определения параметров треков в системе дрейфовых трубок, мы отказались от глобального прослеживания сразу всех треков события и искали треки по данным преобразования Хафа по очереди. При прохождении частицы сквозь дрейфовую трубку регистрируется координата ее центра и расстояние от него до траектории прошедшей частицы (см. рис. 4). Главная неприятность здесь состоит в том, что знание этого расстояния не позволяет определить, слева или справа от центра прошла частица, возникает проблема ***лево-право неопределенности***.

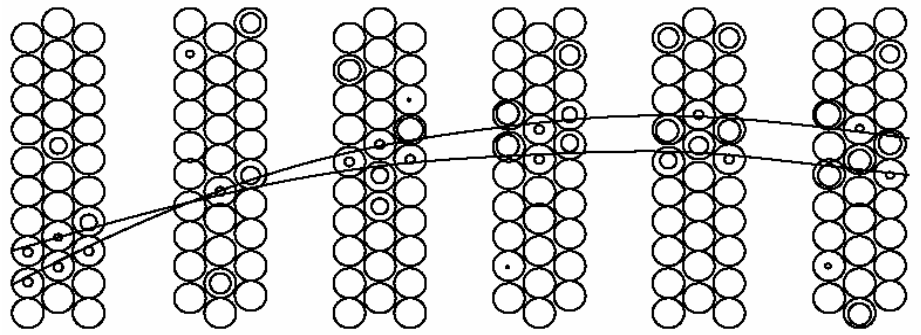

Рис.4. Типичное событие, детектированное дрейфовой камерой в магнитном поле.

Для учета этой двойственности в работе [13] вместо бинарных нейронов S_{ia}, используемых для определения принадлежности точки треку, был введен двумерный вектор-нейрон $\mathbf{S}_i = (s_i^+, s_i^-)$ с допустимыми значениями (1,0), (0,1), (0,0), что увеличило размерность минимизируемой функции и привело к появлению двух факторов Поттса.

Основой схемы минимизации явилась так называемая процедура имитационного отжига, для описания которой мы обратимся к работе [8] Джиласси и Харландера (ДХ). Для решения задачи одновременного поиска и подгонки треков ДХ предложили свой метод эластичного трекинга (ЕТ), исходя из другой идеи, не связанной с ИНС (хотя в [16] они показали, как можно трактовать ЕТ-метод как один из вариантов Хопфилдовой ИНС).

ДХ также использовали идею гибкого шаблона, но описывали его, исходя из физического рассмотрения взаимодействия положительно заряженного шаблона и отрицательно заряженных пространственных точек, измеренных на треке. Чем лучше гибкий шаблон пройдет по точкам, тем меньше будет энергия их взаимодействия. Пусть заряд для шаблона трека распределен с плотностью

$\rho_T(r)$, а заряд множества измеренных точек имеет плотность $\rho(r')$. Вычисляя энергию взаимодействия E между двумя этими зарядами, получаем

$$E = -\int dr'dr\rho_T(r)V(r-r')\rho(r') \longrightarrow min, \tag{4}$$

где V – потенциал, зависящий от расстояния измеренных точек до шаблона. ДХ выбрали потенциал Лоренца

$$V(x,t) = \frac{w^2(t)}{x^2 + w^2(t)} \tag{5}$$

с шириной, зависящей от температуры: $w(t)= a + (b - a)\exp(-t/T)$, где t – температура, T – температурная константа, a – максимальное расстояние, на котором точки еще приписываются к данному шаблону, $b \approx \sigma_{res}$ – точность пространственного разрешения детектора. Очевидно, что $b \ll a$. Учитывая дискретность измерений, получаем вместо интеграла в (4) сумму

$$E(\pi,t) = -\frac{1}{N}\sum_i^N \frac{w^2(t)}{(\vec{x}_i - \vec{r}(\pi,\vec{x}))^2 + w^2(t)} \tag{6}$$

Здесь N – число точек на треке, $\mathbf{x}_i$ и $\mathbf{r}_i$ – i-я точка, измеренная в пространстве, и ее расстояние до шаблона, а $\boldsymbol{\pi}$ как и раньше, вектор параметров геликоиды, описывающей траекторию движения частицы в однородном магнитном поле.

Функционал в формуле (6) зависит от точек только одного трека, хотя как и в предыдущем подходе, можно осуществлять одновременную подгонку сразу всех треков. Тем не менее сами ДХ этого не рекомендуют. Наличие зависимости потенциала от температуры позволяет применять технику симулированного отжига (simulated annealing) [17]: вначале, когда параметры известны только в грубом приближении, берется потенциал настолько широким (высокая температура t), чтобы наверняка захватывать все измеренные точки. При этом минимизируемая функция $E(\boldsymbol{\pi},t)$ выполаживается так, что остается только один минимум (см. рис. 5). Хотя он может быть расположен и несколько в стороне от искомого глобального минимума, но может служить начальным приближением на следующем этапе поиска при понижении температурного параметра. При постепенном понижении t до нуля этот процесс сойдется к минимуму исходной функции.

ET метод в такой постановке был успешно применен в работе [18] для распознавания треков по данным проекционной камеры установки STAR в условиях, предельных по загрузке (до 5 тыс. треков на событие), а также в ряде других экспериментальных исследований [19,20].

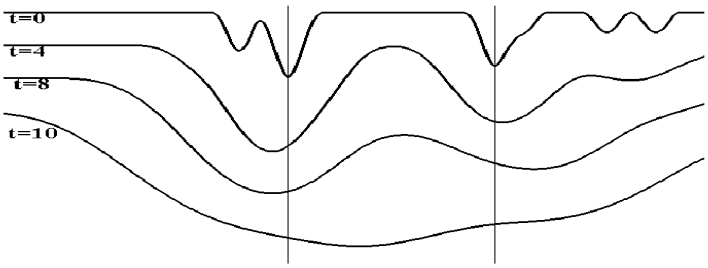

Рис.5. Вид $E(\pi,t)$ при разных температурах.

3. Робастные методы оценки параметров и их применения

Наиболее распространенным методом оценки параметров функциональных зависимостей по данным измерений является метод наименьших квадратов (МНК). Возьмем в качестве простейшего примера линейную модель трека вне магнитного поля в координатной проекции (x,z) : $y_i = y_0 + t_y z_i + \varepsilon_i$; i=1,2, …, n , где y_i – i-е измерение, z_i – координата i-й плоскости детектора, ε_i – ошибка измерения, имеющая по предположению нормальное распределение с нулевым средним и среднеквадратичным значением σ_i. Обозначим вектор неизвестных параметров как $\mathbf{p}^T = (y_0, t_y) = (p_1, p_2)$. Для их определения следует найти минимум суммы квадратов невязок ε_i, которые в общем случае неравноточных измерений должны быть нормированы на их стандартные отклонения σ_i:

$$S(\bar{p}) = \sum_i w_i \epsilon_i^2 \Longrightarrow \min_{\bar{p}}, \tag{7}$$

где $w_i = 1/\sigma_i^2$ – веса измерений. Чтобы найти минимум $S(\mathbf{p})$ приравняем нулю производные этого функционала по параметрам. Это даст систему нормальных МНК-уравнений:

$$\frac{\partial S}{\partial p_j} = 2\sum_i w_i \epsilon_i \frac{\partial \epsilon_i}{\partial p_j} = 0, \quad j = 1,2 \tag{8}$$

решения которой дают искомые оценки параметров. МНК-оценки обладают целым рядом полезных свойств, гарантированных тем, что при непременном условии нормальности распределения невязок ε_i МНК оказывается частным случаем общего метода оценивания, называемого методом максимального правдоподобия (ММП), который действует при произвольных законах распределения оцениваемых параметров. Нарушение нормальности величин ε_i в трековых измерениях обычно является следствием засорения выборки посторонними измерениями как шумовыми, так и от соседних треков и неизбежно ведет к значительным искажениям оцениваемых параметров.

Квадратичность функционала (7) ведет к тому, что точки, далеко отстоящие от подгоняемой кривой, 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 подгоняемой функции, придавая остальным меньшие значения или вообще пренебрегая ими. Такую идею можно реализовать, придавая каждому измерению специальный вес, значение которого убывает с ростом невязки ε_i, т.е. расстояния до подгоняемой кривой. Этот подход, называемый **робастным**[1], был дан П. Хьюбером [21], предложившим, однако, иной метод его реализации. 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 в (7) к сумме некоторых симметричных **функций вклада** $\rho(\varepsilon)$, которые также зависят от отклонения ε точки от прямой, но растут медленнее, чем квадратичная парабола. Теперь минимизируемый функционал будет выглядеть так:

$$L(p) = \sum_i \rho(\varepsilon_i)\,. \qquad (9)$$

Заботясь о выпуклости этого функционала, чтобы обеспечить единственность решения задачи его минимизации, П.Хьюбер предложил неограниченную функцию вклада $\rho(\varepsilon)$, составленную из параболы в малой окрестности нуля, продолженной далее отрезками прямых. Однако это предложение оказалось мало устойчивым к сильным засорениям, так как неограниченная $\rho(\varepsilon)$ по-прежнему придавала неоправданно большой вес точкам, далеко отстоящим от подгоняемой линии. Более популярным оказался весовой подход Дж. Тьюки [22], упрощавший одновременно и вычислительную проблему минимизации функционала (9). Если продифференцировать (9) по параметрам, получим систему нелинейных уравнений

$$\frac{\partial L(p)}{\partial p} = \sum_{i=1}^{n} \frac{\partial \rho(\epsilon_i)}{\partial \epsilon_i} \frac{\partial \epsilon_i}{\partial p} = 0$$

которая в обозначениях

$$w(\epsilon) = \frac{1}{\epsilon} \frac{\partial \rho(\epsilon)}{\partial \epsilon} \qquad (10)$$

превращается в систему

$$\frac{\partial L(p)}{\partial p} = \sum_{i=1}^{n} w(\epsilon_i) \frac{\partial \epsilon_i}{\partial p} \epsilon_i = 0 \qquad (11)$$

[1] robust (англ.) – крепкий, здоровый, в статистике – не чувствительный к шумам.

казалось бы полностью совпадающую с нормальными уравнениями МНК (8), но с тем важным отличием, что числовые весовые коэффициенты заменены на **весовые функции** $w(\varepsilon)$, которые приходится перевычислять на каждой итерации получившейся итеративной процедуры, названной процедурой Флетчера-Гранта-Хеблена (ФГХ) [23]. На каждой итерации ФГХ-процедуры выполняется взвешенный МНК, но с функциональными весами. Если нет каких-либо априорных соображений по выбору начальных значений весов, то можно инициировать ФГХ процедуру с помощью обычного МНК, взяв в качестве весов единицы $w(\varepsilon) \equiv 1$**.**

Важным аспектом ФГХ-процедуры является выбор весовой функции. Из формулы (10) следует, что в обычном МНК весовая функция является единичной константой, в то время как в весовой функции, соответствующей функции вклада Хьюбера, эта константа за пределами малого центрального участка спадает по гиперболе. Отсутствие теоретического обоснования по выбору $w(\varepsilon)$ вызвало массу эвристических предложений (см. обзор и анализ в [24]), среди которых одним из наиболее эффективных с точки зрения шумоподавления оказалась весовая функция Тьюки [22], называемая **бивесовой** за применение биквадрата невязок:

$$w(\varepsilon) = \begin{cases} \left[1 - \left(\dfrac{\varepsilon}{c_T \cdot \sigma}\right)^2\right]^2, & \text{если } |\varepsilon| < c_T \cdot \sigma \\ 0, & \text{в противном случае} \end{cases} \qquad (12)$$

Выбор константы c_T в этой формуле определяет ширину коридора вокруг подгоняемой линии, за пределами которого все точки игнорируются. Использование c_T как температуры в схеме симулированного отжига позволяет справиться с проблемой локальных минимумов при минимизации функционала (9). Явное выражение для оптимальной весовой функции удалось вывести для случая равномерного засорения [25]. Для описания засоренного распределения воспользуемся моделью "больших ошибок" (*gross-error model*) Тьюки [26]:
$f_c(\varepsilon) = (1 - g)\, \varphi(\varepsilon) + g\, h(\varepsilon)$, где $\varphi(\varepsilon)$ – гауссово распределение с нулевым средним и среднеквадратичным значением σ_i. g – параметр засорения, а $h(\varepsilon) = 1/\Delta$ – плотность равномерного распределения в отрезке $(-\Delta/2, \Delta/2)$, где $\sigma \ll \Delta$. Составим логарифмическую функцию правдоподобия для такого распределения, приняв для простоты однопараметрическую зависимость $L(p)$:

$$L(p) = \ln \prod_i f(\varepsilon_i) = \sum_i \ln\left[\frac{1-g}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) + \frac{g}{\Delta}\right].$$

Приравнивая к нулю ее производную по параметру p, мы получим уравнение правдоподобия, которое примет вид (11), если обозначить

$$w_{opt}(\epsilon) = \frac{1+c}{1+c\cdot\exp\left(\frac{\epsilon^2}{2\sigma^2}\right)}, \qquad c = \frac{g}{1-g}\frac{\sqrt{2\pi}\sigma}{\Delta} \tag{13}$$

Как видно, единственный параметр c определяется засоренностью данных не во всем диапазоне измерений Δ , а только в узкой полосе шириной $\sigma\sqrt{2\pi}$ вокруг подгоняемой кривой, где и сосредоточено большинство полезных измерений. Часто константу c можно оценить заранее, исходя из устройства детектора.

Из соображений ускорения вычислений была предложена полиномиальная аппроксимация функции (13) многочленом 4-го порядка, которая оказалась ничем иным, как вышеупомянутым бивесом Тьюки (12).

Уже отмечалась вычислительная проблема минимизации функционала (9), который в силу своей нелинейности может иметь несколько локальных минимумов. Их число зависит не только от оцениваемых параметров p, но и от разброса точек вокруг подгоняемой кривой, т.е. от оценки среднеквадратичного среднего σ. Исследование совместной оценки p и σ было проведено в нашей работе [25] на основе рассмотрения геометрических свойств $L(p,\sigma)$ как функции этих двух параметров. Результаты можно продемонстрировать на простой однопараметрической модели, когда множество локальных условных минимумов $L(p,\sigma)$ для всех фиксированных $\sigma > 0$ образует некий набор гладких кривых в полуплоскости $\{(p, \sigma)$: $\sigma > 0\}$ (см. рис. 6). Обозначим их $\gamma_1, \gamma_2,\ldots, \gamma_m$. Тогда найдется достаточно большое $\sigma_1 > 0$, для которого в полуплоскости $\{(p, \sigma)$: $\sigma > 0\}$ существует только одна из этих кривых, обозначенная как γ_1 . Она неограничена и при $\sigma \to \infty$ асимптотически сходится к среднему всех измеренных значений.

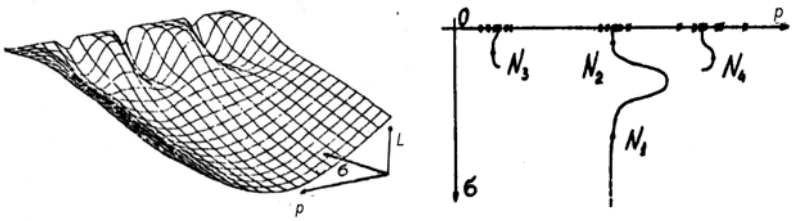

Рис.6. Слева: пример поверхности $L(p,\sigma)$ с характерными оврагами; справа: кривые, соответствующие дну каждого их этих оврагов.

Пример на рис.6 показывает, что возможные случайные сгущения измерений приводят к появлению характерных «оврагов», иногда весьма «кривых», число которых возрастает с уменьшением σ. Детальное теоретическое изучение совместных оценок p и σ в работе [26] и проведенные расчеты на модельных данных показали, что всегда на кривой γ_1 существует по крайней мере одна совместная оценка, являющаяся решением системы уравнений

$$\begin{cases} \partial L/\partial p & = 0; \\ \partial^2 L/\partial p^2 + \sigma^{-1}\partial L/\partial\sigma & = 0. \end{cases} \tag{14}$$

с условиями достижения минимума: $\partial^2 L/\partial p^2 > 0$, $\partial/\partial\sigma\,(\partial^2 L/\partial p^2 + \sigma^{-1}\partial L/\partial\sigma) > 0$.

Если в задаче выполнено условие σ « Δ, то для достаточно большой σ_0, равной, например, размаху выборки x_n - x_1, мы окажемся как раз на кривой γ_1, так что, начиная с асимптотического значения $p_0 = (\sum_i x_i)/n$, мы можем постепенно двигаться вдоль этой кривой до достижения $\sigma = \sigma_{min}$, которая обычно оценивается заранее как среднеквадратичная ошибка ионизированного трека. В реальных вычислениях вместо перевычисления σ^2 на каждой итерации как $\sigma^2=(\sum_i w_i \varepsilon_i^2)/ (\sum_i w_i)$ мы, начиная с σ_0, вычисляем $\sigma^{(k)} =(1-\delta) \sigma^{(k-1)}$ с малым δ=0.05.

4. Приложения и выводы

За последние годы в Объединенном Институте Ядерных Исследований был проведен целый ряд работ по анализу экспериментальных данных на базе применения робастного подхода. Робастные методы подгонки, несмотря на свою кажущуюся простоту, показали высокую эффективность в таких приложениях, как определение вершины взаимодействия с помощью вершинного детектора установки CERES, состоящего всего из двух координатных плоскостей [27], распознавание колец черенковского излучения и проведение идентификации частиц [28,29], поиск треков в экспериментах с высокой множественностью событий [18,30], решение задач калибровки и алайнмента трековых детекторов [31,32]. Отметим развитие теории весовых функций как на случаи неравномерного засорения выборки [30], так и на случаи, когда кроме координат, в детекторе фиксируется и амплитуда сигнала [29] .

Если теперь сравнить процедуры минимизации функционалов (1), (6) и (9) с вычислительной точки зрения, то становится очевидным, что минимизация функционала (1) эластичных нейросетей с помощью итеративного метода (2) сводится, по сути, к взвешенному методу наименьших квадратов с весами (3). Потенциал Лоренца в методе ДХ также может рассматриваться как робастная функция вклада, а вся схема симулированного отжига практически полностью совпадает с вышеописанным методом совместного робастного оценивания параметров положения и масштаба.

Эластичные методы хотя и весьма эффективны в приложениях, но опираются на различные физические аналогии и эвристические обоснования, в то время как робастный подход основан на ясных статистических выводах и по существу всегда может быть сведен к тому или иному эластичному методу как к частному случаю.

Литература

1. Г.А.Ососков, А.Полянский, И.В.Пузынин, Современные методы обработки экспериментальных данных в физике высоких энергий // ЭЧАЯ, т.33, в.3, 2002, 676-745.
2. Р.Стренд, Распознавание оптических образов при экспериментах на трековых камерах с частицами высоких энергий // в сб. «Распознавание образов при помощи цифровых вычислительных машин», под ред. Л.Хармона, пер с англ., М., МИР, 1974,15-37.
3. New computing techniques in physics research III // in: Proc. AIHENP, Ed. K.Becks. Perret-Gallix, World-Sci, Singapore, 1994.

4. И.В.Кисель, В.Н.Нескоромный, Г.А.Ососков, Применение нейронных сетей в экспериментальной физике // ЭЧАЯ, т.24,(1993), вып. 6,1551-1595.
5. L.Gislen, C.Peterson, B.Soderberg, Rotor neurons – basic formalism and dynamics // Neural Computation, 4, 1992,737.
6. G.Ososkov, Robust tracking by cellular automata and neural network with non-local weights // in “Applications and Science of Artificial Neural Networks”, S. K. Rogers, D. W. Ruck, Editors, Proc. SPIE 2492, (1995) 1180-1191.
7. Г.А.Ососков и др., Использование нейронных сетей для улучшения интерпретации эксперимента EXCHARM // Матем. Моделир., т.11, в.10, 1999, 116-126.
8. M.Gyulassy and M.Harlander, Elastic tracking and neural networks algorithms for complex pattern recognition // Comp.Phys.Comm. 66, 1991, 31-46.
9. S.Baginyan et al, Tracking by a Modified Rotor Model of Neural Network // Comp. Phys. Commun. (1994) v.79, 165-178.
10. Galkin et al, Feedback neural networks for ARTIST ionogram processing // Radio Science v.31, № 5 (1996) 1119-1128.
11. M. Ohlsson, C. Peterson, A. Yuille, Track finding with deformable templates - the elastic arms approach // Comput. Phys. Commun., 71, (1992), 77.
12. L. Muresan, R. Muresan, G. Ososkov, Yu. Panebratsev, Deformable Templates for Circle Recognition // JINR Rapid Communications, 1[81]-97, Dubna, 1997, 27-44.
13. S.Baginyan, G.Ososkov, Finding tracks detected by a drift tube system // Comp.Phys.Comm, v. 108, No 1 (1998) 20-28.
14. G.Ososkov, V.Palichik, E.Tikhonenko, Robust Technique with Sub-Optimal Weight Function for Track Fitting in CMS Muon Strip Chamber // Abst. of Europhysics Conf. on Computat. Phys., Vol. 22F, EPS, Granada, Spain (1998) 323-324.
15. Hough P. V. C. A Method and Means for Recognizing Complex Patterns //, US Patent: 3,069,654, 1962.
16. Toft P. The Radon Transform. Theory and Implementation // Ph.D. Thesis, Department of Mathematical Modelling, Section for Digital Signal Processing, Technical University of Denmark, 1996. (см. http://www.ei.dtu.dk/staff/ptoft/ptoft.html)
17. S.Kirkpatrick et al, Optimization by simulated annealing // Science 220 (1983), 671.
18. B.Lasiuk, D.Lyons, G.Ososkov, T.Ullrich, Development of an Elastic Tracking Package // Proc. of CHEP'98, Chicago (1998).
19. R.Blankenbecler, A unified treatment of track reconstruction and particle identification // Comput. Phys. Commun., 81, 1994, 335-342.
20. M.Lindstrom, Track reconstruction in the STLAS detector using elastic arms // Nucl. Instr. Meth. A357, 1995, 129-149.
21. P.Huber, Robust statistics // Wiley, N-Y (1981).
22. F.Mosteller, W.Tukey, Data analysis and regression: a second course in statistics, //Addison - Wesley, N-Y (1977).
23. R.Fletcher et al // Comp.J. v.72, No 3, (1971) p 276.
24. А.Астапов и др., Численный анализ робастных регрессионных методов // Сообщ. ОИЯИ Р9-85-492, Дубна, 1985.
25. G.Ososkov, Robust regression for the heavy contaminated sample // Proc. 2-nd International Tampere Conference in Statistics, (Tampere, Finland) (1987), 615-626.
26. J.Tukey, Introduction to Today’s Data Analysis, Critical Evaluation of Chemical and Physical. Structural Information // Nat. Acad. Sci., Washington, 1974, 3-14.
27. H.Agakishiev et al, New Robust Fitting Algorithm For Vertex Reconstruction in the CERES Experiment // Nucl. Instr. and Methods, A394 (1997), 225-231.
28. G.Agakichiev et al, Cherenkov Ring Fitting Techniques for the CERES RICH Detectors // Nucl. Instr. and Methods, A371 (1996) 243-247.
29. N.Chernov, E.Kolganova, G.Ososkov, Robust Methods for the RICH Ring Recognition and Particle Identification // Nuclear Inst. Meth. A433 (1999) 274-278.

30. Golutvin, Yu.Kiryushin, S.Movchan, G.Ososkov, V.Palichik, E.Tikhonenko, Robust estimates of track parameters and spatial resolution CMS muon chambers // Comp.Phys.Comm, v.126/1-2, (2000) 72-76.
31. A.Bel'kov, A.Moshkin, G.Ososkov, Autocalibration of PC OTR chambers based on the robust fitting approach // HERA-B Note 02-067, DESY, Hamburg, 2002, 37 pp.
32. O.Barannikova et al, Specifications of SVT global alignment package, STAR Note 0364, BNL, 1998.

Статья поступила 12 декабря 2005 г.