*ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И АТОМНОГО ЯДРА*
*2002. Т. 33. ВЫП. 3*

УДК 519.254; 519.237

# СОВРЕМЕННЫЕ МЕТОДЫ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ В ФИЗИКЕ ВЫСОКИХ ЭНЕРГИЙ

*Г. А. Ососков, А. Полянский, И. В. Пузынин*

Объединенный институт ядерных исследований, Дубна
Лаборатория информационных технологий

*ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И АТОМНОГО ЯДРА*
*2002. Т. 33. ВЫП. 3*

УДК 519.254; 519.237

# СОВРЕМЕННЫЕ МЕТОДЫ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ В ФИЗИКЕ ВЫСОКИХ ЭНЕРГИЙ

*Г. А. Ососков, А. Полянский, И. В. Пузынин*

Объединенный институт ядерных исследований, Дубна
Лаборатория информационных технологий

Обсуждаются три основных метода обработки экспериментальных данных, активно используемых в последние годы в Объединенном институте ядерных исследований: робастные методы математической статистики, искусственные нейронные сети, клеточные автоматы и вейвлет-анализ. Обзор сделан, главным образом, по работам, выполненным с участием сотрудников Лаборатории информационных технологий, в том числе и в рамках международных коллабораций с крупными физическими центрами: CERN, DESY, BNL и др. Авторы постарались достаточно подробно осветить основные понятия обсуждаемых методов и привести наиболее полезные и перспективные примеры их применения.

In the given survey three basic methods of experimental data processing are considered which are intensively used during last decade at the Joint Institute for Nuclear Research, namely: robust methods of mathematical statistics, artificial neural networks, cellular automata and wavelet analysis. The main source of survied papers is those that have been elaborated by authors from Laboratory of Informational Technologies, as well as works developed in collaborations with famous physical centres as CERN, DESY, BNL, etc. The authors described the basic principles of discussed methods and gave most useful and promising examples of their applications.

## ВВЕДЕНИЕ

Достижения теоретической физики в последние десятилетия вызвали настоящий переворот в экспериментальной физике высоких энергий (ФВЭ), перешедшей к коллайдерным ускорителям на встречных пучках в тэвных областях энергий, таким как уже действующий RHIC (BNL) или почти готовый к запуску LHC в CERN. Соответственно кардинально сменились техника и технология детектирующих комплексов, использующих последние открытия в области электронных технологий, распределенных вычислительных систем и результаты в программировании, связанные с переходом к объектно-ориентированной структуре программ и взрывным распространением Интернета. Ряд действующих и проектируемых крупных экспериментов в ФВЭ, проводимых в настоящее время в ведущих мировых исследовательских институтах ядерной физики, выполняются в сотрудничестве с Объединенным

институтом ядерных исследований в Дубне. Можно упомянуть такие эксперименты, как NA45/CERES, COMPASS, CMS, ATLAS (CERN), STAR (BNL). Они направлены на исследование чрезвычайно редких явлений, предсказанных современными физическими теориями: нарушение электрослабой симметрии, кварковые и лептонные структуры и т. д. В состав этих экспериментов входят современные электронные детекторы: пропорциональные, времяпроекционные, силиконовые камеры, дрейфовые трубки, RICH-детекторы колец черенковского излучения и различные калориметры. Определяющими признаками данных, поступающих с этих детекторов, являются:

• дискретный характер и сложность текстуры распознаваемых образов; высочайшие темпы поступления данных;

• превышение на несколько порядков числа фоновых событий над числом полезных в условиях, когда первые по своему характеру близки ко вторым;

• высокий уровень и коррелированность шумовых отсчетов;

• большая множественность объектов (следов частиц: треков, колец черенковского излучения, ливней), подлежащих распознаванию в каждом событии;

• необходимость в сложном многопараметрическом преобразовании дискретных данных детекторов в стандартную систему координат. Поскольку каждый детектор является системой, собираемой из многих компонент, их взаимное расположение может быть нарушено при сборке, что ведет к систематическому искажению регистрируемых данных. Определение и компенсация этих искажений математическим путем (эта процедура называется «алайнмент» (alignment — дословно: выравнивание)) — обязательное условие правильной обработки, чрезвычайно осложняемой принципиальной невозможностью иметь какой-либо эталонный объект для фиксации возможных искажений.

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 вышеперечисленных условий 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. Таким образом, возникшая насущная потребность в разработке нового математического и алгоритмического аппарата потребовала привлечения математических средств, как новых, так и хотя известных, но мало использовавшихся из-за слабости старой компьютерной базы. Сюда относится применение следующих математических методов:

• робастных методов математической статистики для быстрой подгонки пространственных траекторий частиц, определения первичных и вторичных вершин и разделения близких колец черенковского излучения;

• искусственных нейронных сетей (ИНС) и клеточных автоматов (КА) как для распознавания треков заряженных частиц, так и для проверки физических гипотез, а также для обработки изображений;

• преобразований Радона–Хафа, а также методов, обеспечивающих «сверхразрешение» перекрывающихся сигналов для вычисления начальных значений параметров в итерационных процедурах;

• вейвлет-преобразований (wavelet-transform) для быстрого выделения признаков, разделения близких сигналов и сглаживания данных;

• быстрых вычислительных алгоритмов поиска глобальных экстремумов нелинейных целевых функционалов, а также комбинированных, гибридных методов, обеспечивающих высокую эффективность процессов обработки.

В обзоре рассматриваются работы, выполненные сотрудниками ОИЯИ, по разработке и применению вышеперечисленных новых методов компьютерной обработки данных современных экспериментов по ФВЭ.

В разделе 1 описаны принципы робастных $M$-оценок параметров, сводящихся к итеративному взвешенному методу наименьших квадратов, дан вывод оптимальных весовых функций и приведен богатый набор примеров успешных применений робастного подхода к оценкам параметров в условиях зашумленности данных, в том числе и для задач алайнмента.

Раздел 2 посвящен описанию различных применений ИНС и клеточных автоматов, в том числе нового типа ИНС для обработки изображений.

В разделе 3 дано описание так называемых эластичных методов трекинга (называемых также методом деформируемых шаблонов), нескольких его применений, а также его связи с робастными методами.

В разделе 4 приведено краткое описанию вейвлет-преобразований и рассмотрено их применение для анализа экспериментальных данных.

В заключении подведены краткие итоги описанных в обзоре современных методов анализа данных и указаны некоторые тенденции их развития.

## 1. РОБАСТНЫЕ МЕТОДЫ ОЦЕНКИ ПАРАМЕТРОВ И ИХ ПРИМЕНЕНИЯ

**1.1. Метод наименьших квадратов и необходимость робастного подхода.** Наиболее распространенным методом оценки параметров функциональных зависимостей по данным измерений является метод наименьших квадратов (МНК). Возьмем в качестве примера простую линейную модель трека

вне магнитного поля в координатной проекции $(y, z)$

$$y_i = y_0 + t_y z_i + \epsilon_i, \quad i = \overline{1, n}, \tag{1}$$

где $y_i$ — $i$-е измерение; $z_i$ — координата $i$-й плоскости детектора; $\epsilon_i$ — ошибка измерения, имеющая по предположению нормальное распределение

$$\epsilon_i \in \mathcal{N}(0, \sigma_i), \tag{2}$$

а $\mathbf{p}^T = (y_0, t_y) = (p_1, p_2)$ — вектор неизвестных параметров. Для их определения следует найти минимум суммы квадратов невязок $\epsilon_i$, которые в общем случае неравноточных измерений должны быть нормированы на их стандартные отклонения $\sigma_i$:

$$S(\mathbf{p}) = \sum_i w_i \epsilon_i^2 \Longrightarrow \min_{\bar{p}}, \tag{3}$$

где $w_i = 1/\sigma_i^2$ — веса измерений.

Чтобы найти минимум $S(\mathbf{p})$, приравняем нулю производные этого функционала по параметрам $\mathbf{p}$. Это даст систему нормальных МНК-уравнений:

$$\frac{\partial S}{\partial p_j} = 2 \sum_i w_i \epsilon_i \frac{\partial \epsilon_i}{\partial p_j} = 0, \quad j = 1, 2, \tag{4}$$

решения которой $\hat{\mathbf{p}}^T = \{\hat{p}_j\}$ дают искомые оценки параметров. Значения функционала (3) в минимуме, т. е. при найденных $\{\hat{p}_j\}$, служат количественной оценкой качества подгонки, т. к. величина

$$S(\hat{\mathbf{p}}) = \sum_i^n \left( \frac{y_i - \hat{p}_1 - \hat{p}_2 z_i}{\sigma_i} \right)^2 \tag{5}$$

оказывается распределенной по закону $\chi^2$ с $m - 2$ степенями свободы [1].

В более общем случае линейной регрессии, например подгонки полинома $y = \sum_{j=1}^{m+1} p_j z^{(j-1)}$ с $m + 1$ параметрами, удобнее перейти к матричным обозначениям:

$$\mathcal{Y} = \mathcal{Z}\mathcal{P}, \tag{6}$$

где

$$\mathcal{Z} = \begin{pmatrix} 1 & z_1 & z_1^2 & \dots & z_1^m \\ 1 & z_2 & z_2^2 & \dots & z_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_n & z_n^2 & \dots & z_n^m \end{pmatrix}, \ \mathcal{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathcal{P} = \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_m \end{pmatrix}. \tag{7}$$

Определяя также весовую матрицу

$$W = \begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{pmatrix}, \tag{8}$$

можно переписать выражение (3) в матричной форме

$$\mathcal{S} = \mathcal{E}^T W \mathcal{E}, \tag{9}$$

где

$$\mathcal{E} = \mathcal{Y} - \mathcal{Z}\mathcal{P}. \tag{10}$$

Тогда решением нормальной системы уравнений становится

$$\hat{\mathcal{P}} = (\mathcal{Z}^T W \mathcal{Z})^{-1} \mathcal{Z}^T W \mathcal{Y}. \tag{11}$$

Это дает нам оценки всех $m+1$ параметров. Их ковариационная матрица

$$\mathrm{cov}_{\hat{\mathbf{p}}} = \mathcal{C} = (\mathcal{X}^T W \mathcal{X})^{-1}. \tag{12}$$

Дисперсии параметров расположены по диагонали, но поскольку сами параметры оказываются зависимыми, их ковариации уже не равны нулю. Ковариация параметров $p_i$ и $p_j$ равна элементу $c_{ij}$ этой матрицы.

Популярность МНК среди экспериментаторов объясняется тем, что получаемые МНК-оценки параметров оказываются состоятельными и асимптотически эффективными, т. е. наилучшими среди всех других методов конструирования оценок (см., например, [1]). Это следует из того, что при сделанном нами предположении (2) МНК оказывается частным случаем общего метода оценивания, называемого методом максимального правдоподобия (ММП), который действует при произвольных законах распределения оцениваемых параметров. В предположении, что все невязки (10) являются независимыми случайными величинами с нулевыми средними и одинаковой функцией плотности распределения (ФПР) $f(e)$, вероятность появления конкретной выборки $e_1, e_2, \ldots e_n$ является *функцией правдоподобия*

$$L(\mathbf{p}) = \prod_i^n f(e_i). \tag{13}$$

Согласно принципу *максимального правдоподобия* Р. Фишера, наилучшей — наиболее правдоподобной — оценкой параметров будет тот набор параметров $\mathbf{p}$, при котором (13) примет свое максимальное значение. Поскольку и сама

функция правдоподобия, и ее логарифм принимают максимальное значение в одной и той же точке, соответствующее *уравнение правдоподобия* для ее поиска записывается так:

$$\frac{\partial \ln L(\mathbf{p})}{\partial p_k} = 0; \quad k = 1, 2, \ldots, m. \tag{14}$$

Решение этого уравнения и дает нам ММП-оценки $\hat{\mathbf{p}}$, которые, как доказано [1], обладают свойствами состоятельности, асимптотической нормальности и эффективности.

Поэтому предположение о нормальности ФПР в (13), т. е.

$$f(e_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right), \tag{15}$$

является весьма важным, так как только в этом случае ММП превращается в МНК (с сохранением всех полезных свойств получаемых оценок). В самом деле,

$$\ln L(\mathbf{p}) = \ln\left(\prod_{i=1}^{n} f(e_i)\right) = \sum_{i=1}^{n} \ln f(e_i),$$

и логарифмическая функция правдоподобия оказывается равной

$$-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{e_i^2}{\sigma^2}\right) + \text{const}.$$

Она имеет свой максимум там же, где (13) имеет свой минимум.

К сожалению, ключевое предположение о нормальности невязок в реальной жизни нарушается очень часто благодаря засорению измерений шумовыми или фоновыми измерениями, как, например, показано на рис. 1.

Квадратичность функционала $S$ в (9) ведет 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 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 $e_i$, т. е. расстояния до подгоняемой кривой. Этот подход, называемый *робастным**, был дан П. Хьюбером [2], предложившим, однако, иной метод его реализации. Предложение Хьюбера сводится к некоторому обобщению метода максимального

*Robust (англ.) — крепкий, здоровый. В статистике — нечувствительный к шумам.

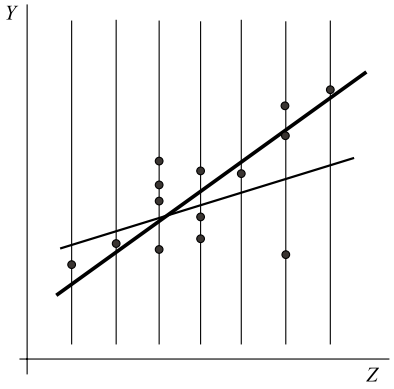



Рис. 1. Пример данных трековых измерений при наличии шумовых срабатываний детектора и фоновых измерений: толстая линия — трек; тонкая — результат МНК-подгонки

правдоподобия. Подчеркивая эту связь с ММП, Хьюбер назвал свой подход *М-оцениванием*. С математической точки зрения предлагалось перейти от суммы квадратов в (3) к сумме некоторых *функций вклада* $\rho(e)$, которые также зависят от отклонения $e$ точки от прямой, но растут медленнее, чем квадратичная парабола.

Теперь минимизируемый функционал будет выглядеть так:

$$L(p) = \sum_{i=1}^{n} \rho(e_i). \tag{16}$$

Для простоты мы здесь рассматриваем функцию $L$ как зависящую от одного параметра. Так как точки, близкие к прямой, вероятнее всего, ей принадлежат, функция $\rho(e)$ при малых $e$ должна себя вести как $e^2$, как и в методе наименьших квадратов:

$$\rho(e) \sim e^2 \text{ при } e \to 0. \tag{17}$$

Исходя из этого условия и требования единственности решения, что может быть обеспечено выпуклостью функционала (16), Хьюбер предложил выпуклую функцию вклада (см. на рис. 2, *б*)

$$\rho(e) = \begin{cases} e^2, & |e| < e_0 \\ 2e_0|e| - e_0^2, & |e| > e_0, \end{cases} \tag{18}$$

которая при больших отклонениях, $|e| > e_0$, возрастает по линейному закону. При использовании такой функции вклада точки, лежащие далеко от прямой, меньше влияют на подгонку, чем в случае квадратичной зависимости, хотя и продолжают влиять в силу неограниченности функции вклада.

В этой связи другие исследователи предложили использовать ограниченные функции вклада, вообще игнорирующие точки за пределами некой полосы вокруг подгоняемой кривой. Простейшим примером такой ограниченной функции вклада может служить «обрезанная» сверху парабола (см. рис. 2, *в*):

$$\rho(d) = \begin{cases} d^2, & |d| < d_0 \\ d_0^2, & |d| > d_0. \end{cases} \tag{19}$$

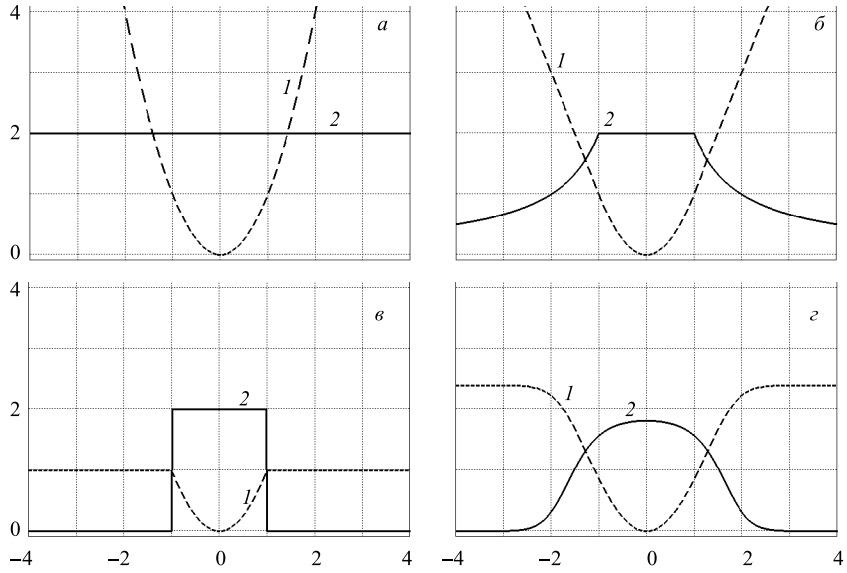



Рис. 2. Неограниченные (*а*, *б*) и ограниченные (*в*, *г*) функции вклада (кривые *1*) и соответствующие им весовые функции (кривые *2*): *а*) МНК-парабола; *б*) функция вклада Хьюбера; *в*) «обрезанная» парабола; *г*) оптимальная функция вклада

Недостатком такой функции являются ее негладкость, создающая проблемы при дифференцировании $L(\mathbf{p})$ во время решения уравнения правдоподобия для нахождения минимума. Общим же недостатком ограниченных функций вклада является то, что их использование ведет к невыпуклости функционала (16), приводящей к появлению нескольких минимумов $L(\mathbf{p})$, так что поиск среди них глобального может оказаться серьезной вычислительной проблемой.

Говоря о вычислительной стороне робастных методов, следует отметить, что платой за робастность оказывается возникающая нелинейность задачи, даже в случаях подгонки функций, линейно зависящих от параметров. Таким образом, робастные методы оказываются итеративными по самой своей природе.

**1.2. Математический формализм робастного подхода.** Рассмотрим уравнение правдоподобия для функционала (16)

$$\frac{\partial L(p)}{\partial p} = \sum_{i=1}^{n} \frac{\partial \rho(e_i)}{\partial e_i} \frac{\partial e_i}{\partial p} = 0 \tag{20}$$

и преобразуем его, обозначив

$$w(e) = \frac{1}{e}\frac{\partial \rho(e)}{\partial e}, \tag{21}$$

$$\frac{\partial L(p)}{\partial p} = \sum_{i=1}^{n} w(e_i)\frac{\partial e_i}{\partial p}e_i = 0. \tag{22}$$

Мы получили уравнение, аналогичное нормальным уравнениям МНК (4), но с заменой числовых весовых коэффициентов на *весовые функции* $w(e)$, которые приходится заново вычислять на каждой итерации получившейся итеративной процедуры, названной процедурой Флетчера–Гранта–Хеблена (ФГХ) [4]. 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 $w_i^{(0)} \equiv 1$.

Одной из наиболее эффективных с точки зрения подавления шумов является весовая функция Тьюки [5], называемая *бивесовой* за применение биквадрата невязок:

$$w(e) = \begin{cases} \left(1 - \left(\dfrac{e}{c_T\sigma}\right)^2\right)^2, & \text{если } |e| < c_T\sigma, \\ 0 & \text{в остальных случаях.} \end{cases} \tag{23}$$

Некоторые из вышеупомянутых функций вклада и соответствующие им весовые функции изображены на рис. 2.

Можно заметить, что для метода наименьших квадратов каждая точка имеет одинаковый вес, при использовании функции вклада Хьюбера вес точек, лежащих далеко от траектории, уменьшается обратно пропорционально отклонению от траектории, а ограниченная функция вклада приводит к тому, что далекие точки вообще не учитываются.

Сравнительный анализ многих вариантов ограниченных и неограниченных функций вклада можно найти в [3].

Многообразие имеющихся предложений по выбору весовых функций стимулировало исследование [6] по определению весовой функции, оптимальной при подгонке кривой на фоне равномерного засорения. Будем по-прежнему предполагать, что измерения производятся с ошибками, распределенными по нормальному закону (15) со стандартным отклонением $\sigma$, а засоряющие точки распределены равномерно на гораздо более широком интервале $\Delta\,(\sigma \ll \Delta)$. Предположим также, что отношение сигнал/шум равно $(1-\varepsilon)/\varepsilon$, т. е. $\varepsilon$ — это параметр засорения.

Для описания засоренного распределения воспользуемся моделью «больших ошибок» (*gross-error model*) Тьюки

$$f_\varepsilon(e) = (1-\varepsilon)\phi(e) + \varepsilon h(e), \tag{24}$$

где $\phi(e)$ — гауссово распределение (15), а $h = 1/\Delta$ — плотность равномерного распределения.

Составим логарифмическую функцию правдоподобия для такого распределения, рассматривая по-прежнему для простоты однопараметрическую зависимость

$$L(p) = \ln \prod_i f(e_i) = \sum_i \ln \left( \frac{1-\varepsilon}{\sqrt{2\pi}\sigma} \, \mathrm{e}^{-\frac{e^2}{2\sigma^2}} + \frac{\varepsilon}{\Delta} \right). \tag{25}$$

Приравнивая к нулю ее производную по параметру $p$, мы получим уравнение правдоподобия

$$\frac{\partial \ln L}{\partial p} = \sum_i \frac{\dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma} \, \mathrm{e}^{-\frac{e^2}{2\sigma^2}} e_i \dfrac{\partial e_i}{\partial p}}{\dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma} \, \mathrm{e}^{-\frac{d^2}{2\sigma^2}} + \dfrac{\varepsilon}{\Delta}} = 0, \tag{26}$$

которое может быть переписано в виде, аналогичном (4):

$$\frac{\partial \ln L}{\partial p} = \sum_i w(e_i) e_i \frac{\partial e_i}{\partial p} = 0,$$

где обозначено

$$w(e) = \frac{\dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma} \, \mathrm{e}^{-\frac{e^2}{2\sigma^2}}}{\dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma} \, \mathrm{e}^{-\frac{d^2}{2\sigma^2}} + \dfrac{\varepsilon}{\Delta}}.$$

Таким образом, после деления на числитель и нормировки на единицу в нуле мы получим выражение для оптимальных весов:

$$w_{\text{opt}}(e) = \frac{1+c}{1 + c \, \exp\left( \dfrac{e^2}{2\sigma^2} \right)}, \tag{27}$$

где $c = \dfrac{\varepsilon}{1-\varepsilon} \dfrac{\sqrt{2\pi}\sigma}{\Delta}$.

Поскольку $\Delta \gg \sigma$, эта константа $c$ довольно мала даже при малом отношении сигнал/шум. Оптимальная весовая функция и соответствующая ей

функция вклада изображены на рис. 2, *г*. Из соображений ускорения вычислений была предложена полиномиальная аппроксимация функции (27) многочленом 4-го порядка, которая оказалась ничем иным, как вышеупомянутым бивесом Тьюки (23). В этой связи бивесовая функция используется весьма часто, хотя следует подчеркнуть необходимость тщательного выбора параметра $c_T$. Есть рекомендации [5] $c_T = 3 \div 5$. При этом следует учитывать, что выброс измерений с большими отклонениями путем приписывания им нулевых весов идет по величине $c_T\sigma$, где значение $\sigma$ также подлежит перевычислению на каждой итерации:

$$\sigma^{(k)} = \sqrt{\sum_i w_i^{(k-1)} \left(e_i^{(k-1)}\right)^2 \Big/ \sum_i w_i^{(k-1)}}. \qquad (28)$$

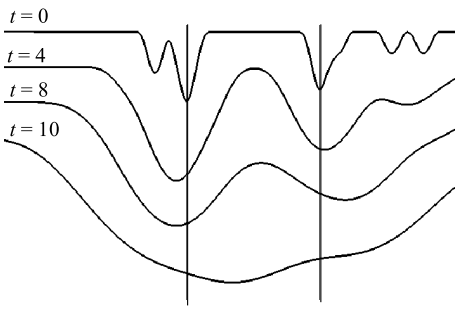



Рис. 3. Вид функции правдоподобия после введения температурной зависимости при разных температурах

Гибкость робастной процедуры минимизации проявляется особенно значимо в тех случаях, когда процесс минимизации застревает из-за скатывания функционала в локальный минимум. Известным подходом для преодоления подобных трудностей является метод «имитации отжига» (simulated annealing) [7], состоящий в растяжении, выполаживании минимизируемого функционала путем введения некоторого параметра $t$, имитирующего повышение «температуры» так, чтобы локальные минимумы сгладились и остался только один (см. рис. 3). Хотя он и может быть расположен несколько в стороне от искомого глобального минимума, но служит начальным приближением на следующем этапе поиска при понижении температурного параметра. При постепенном понижении $t$ до нуля этот процесс сойдется к минимуму исходного функционала. Точно такой же процесс может быть реализован и в робастной схеме путем введения аналогичной температурной зависимости параметра $c_T$, например, как $c_T = 3 + t$ при $t = 10, 8, ..., 0$.

*Замечание 1.* Следует, конечно, помнить, что применение робастного $M$-оценивания параметров имеет смысл только в случаях засоренности выборки. Если, напротив, есть уверенность, что выборка свободна от фоновых измерений, то никакой необходимости в использовании итеративной процедуры с функциональными весами нет, т. к. для таких случаев доказана максимальная эффективность обычного МНК.

*Замечание 2.* Существенно, что вышеприведенный вывод весовых функций сделан на основе только учета распределения невязок, независимо от принятой параметризации. Поэтому общее решение (11) регрессионной задачи в матричных обозначениях (6)–(11) остается справедливым на каждой ФГХ-итерации, надо только заменить постоянные веса в весовой матрице (8) на их функциональные выражения (27) или (23).

В заключение этого формально-аналитического раздела перед переходом к описанию физических приложений укажем кратко на важное пространственное обобщение робастного подхода, порожденное как раз прикладными проблемами, возникшими благодаря развитию детекторных технологий. Речь идет о новых типах детекторов, обладающих высоким пространственным разрешением, достигнутым повышением их гранулярности: проходящая частица регистрируется не просто как точка на одной из координатных плоскостей детектора, а как кластер из смежных ячеек, между которыми распределилась энергия пролетевшей частицы. Обычно такое распределение имеет колоколообразный вид, хорошо аппроксимируемый двумерным гауссовым распределением типа

$$g(x, y, ; x_0, y_0) = A \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2}\right), \tag{29}$$

где $(x_0, y_0)$ — центр кластера, т. е. место прохождения частицы через данную плоскость детектора; $A$ — амплитуда сигнала, связанная с суммарной энергией, рассеянной на этой плоскости. При детектировании происходит оцифровка сигналов, так что часть экспериментальных данных, относящихся к данной координатной плоскости, выглядит как двумерная гистограмма. Такая картина характерна, например, для детекторов типа силиконовых дрейфовых или времяпроекционных камер, состоящих из нескольких таких координатных плоскостей, а также и для некоторых высокогранулярных детекторов черенковского излучения типа CERES RICH [9, 95].

При этом в силу многократного рассеяния координаты $(x_0, y_0)$ на каждой из них получают случайные добавки, кроме того, содержимое каждой такой ячейки гистограммы $a_{ij}$, получаемое путем интегрирования гауссиана (29) по площади ячейки $ij$, также получает случайные добавки вследствие ошибок измерения.

Таким образом, вместо известной задачи пространственной подгонки треков по точкам мы получаем совершенно новую задачу подгонки по множеству точек — середин ячеек гистограмм $(\bar{x}_i, \bar{y}_i)$, да еще снабженных амплитудами $a_{ij}$. Физики сводят эту задачу к известной подгонке по точкам, которые получают из «сырых» данных путем их специальной предобработки: класте-

ризации и затем определения центра тяжести каждого кластера по формуле

$$x_{\mathrm{cog}} = \left(\sum_i a_{ij}\bar{x}_i\right) \Big/ \left(\sum_i a_{ij}\right); \quad y_{\mathrm{cog}} = \left(\sum_j a_{ij}\bar{y}_j\right) \Big/ \left(\sum_j a_{ij}\right), \quad (30)$$

где $\bar{x}_i = \underset{z}{(x_{i+1} + x_i)}$, $\bar{y}_i = \underset{z}{(y_{j+1} + y_j)}$ — координаты середин ячеек.

Робастный подход позволяет осуществить подгонку непосредственно по «сырым» данным, минуя этап препроцессинга и учитывая дополнительную экспериментальную информацию, содержащуюся в данных об амплитудах зарегистрированных сигналов. Подгонка выполняется итеративным взвешенным МНК по всему множеству данных, включая все центры ячеек и их амплитуды. Несложно найти вес любой из ячеек. На каждой координатной плоскости вычисляется расстояние $e_{ij}$ от центра каждой ячейки до подгоняемой кривой, после чего ее вес вычисляется по формулам (27) или (23).

Гораздо сложнее учесть влияние на этот вес ее амплитуды $a_{ij}$. Простейшее решение: вес, вычисленный, скажем, по формуле (23), умножается на амплитуду $a_{ij}$, соответствующую данной ячейке. Однако при этом не учитывается распределение этих амплитуд, которое может быть, например, экспоненциальным с очень большим разбросом значений. Из-за этого вероятность того, что посторонняя точка, лежащая в стороне, получит неоправданно большой вес за счет своей большой амплитуды, может оказаться слишком велика. И наоборот, точка с малой амплитудой, лежащая чуть в стороне от трека, там, где она и должна быть, ошибочно получит слишком малый вес. Зависимость весовой функции от амплитуды должна быть двумерной: если амплитуда велика, то должно быть более правдоподобно, что такая точка близка к треку. Точки же с малой амплитудой, скорее всего, должны лежать на некотором удалении либо слева, либо справа от трека.

В работе [8] это качественное рассуждение удалось реализовать в виде строгого математического вывода. В обозначениях $f(a)$ — для ФПР амплитуды полезного сигнала $A$, $u(A)$ — для ФПР фоновых амплитуд, в предположении гауссовой формы сигнала вида (29) с $\sigma_x = \sigma_y = \sigma_0$ была получена формула для двумерной весовой функции

$$w(e; A) = \frac{g\, f(gA) + g^2 A\, f'(gA)}{g\, f(gA) + u(A)}, \quad (31)$$

где $g(e) = \sqrt{2\pi\sigma_0^2}\, \mathrm{e}^{e^2/2\sigma_0^2}$. Если взять распределение $f(a)$ показательным со средним $A_0$, т. е. $f(a) = A_0^{-1}\, \mathrm{e}^{-a/A_0}, a > 0$, то для широкого выбора фоновой составляющей $u(A)$ (равномерной или гауссовой с большой $\sigma_{\mathrm{noise}}$) мы получим необычную «двурогую» поверхность, кусочно-линейная аппроксимация которой представлена на рис. 4.

**1.3. Применение робастных методов в физике частиц.** Как уже отмечалось выше, чрезвычайно высокая множественность событий, характерная для современных экспериментов, приводит к необходимости анализа данных в условиях, экстремальных как по плотности анализируемых объектов (треков, колец черенковского излучения), так и по фоновой загрузке и зашумленности. В таких условиях любые модификации известных МНК-процедур с выбросом далеко отстоящих точек либо не работают, либо не позволяют добиться требуемой точности. В это же время робастные методы подгонки, несмотря на свою кажущуюся простоту, показали высокую эффективность в самых разных приложениях.

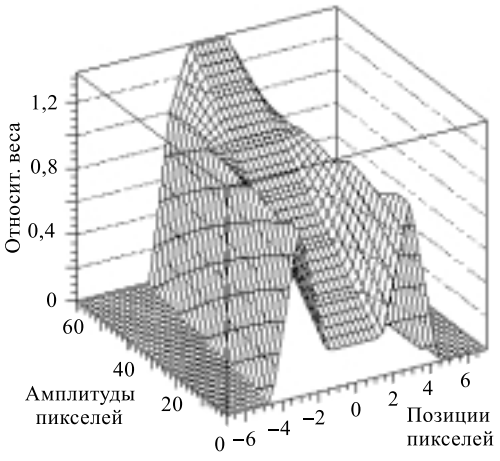



Рис. 4. Двумерная весовая функция, зависящая как от расстояния точки до подгоняемой кривой, так и от амплитуды сигнала в этой точке

*1.3.1. Определение вершины взаимодействия при малом числе координатных плоскостей.* Вершинный детектор установки CERES, состоящий всего из двух силиконовых дрейфовых камер ($\mathrm{SDD}_1$ и $\mathrm{SDD}_2$) [14], предназначен для определения координаты вершины взаимодействия, которое может произойти в любом из восьми дисков мишени (см. рис. 5). Наличие в среднем 700 треков на событие в узком угловом аксептансе на фоне значительного числа шумовых отсчетов делает невозможным предварительное распознавание отдельных треков. Тем не менее задача была успешно решена в [14] путем минимизации функционала вида

$$L(x_v, y_v, z_v) = \sum w_i {e_i}^2, \tag{32}$$

где $x_v, y_v, z_v$ — искомые координаты вершины, а невязка $e_i$ определяется как расстояние от вершины до прямой, соединяющей точки $(x_{i1}, y_{i1})$, $(x_{i2}, y_{i2})$, выбранные на $\mathrm{SDD}_1$ и $\mathrm{SDD}_2$ так, чтобы удовлетворять условиям попадания в угловой аксептанс:

$$e_i = \sqrt{\left(x_v - x_{i1} - \frac{z_v - z_{i1}}{z_{i2} - z_{i1}}(x_{i2} - x_{i1})\right)^2 + \left(y_v - y_{i1} - \frac{z_v - z_{i1}}{z_{i2} - z_{i1}}(y_{i2} - y_{i1})\right)^2}.$$

Применялась бивесовая функция (23) с $c_T = 3$. Итерационная ФГХ-процедура сходилась в среднем за 5 итераций, хотя начальное приближение выбиралось

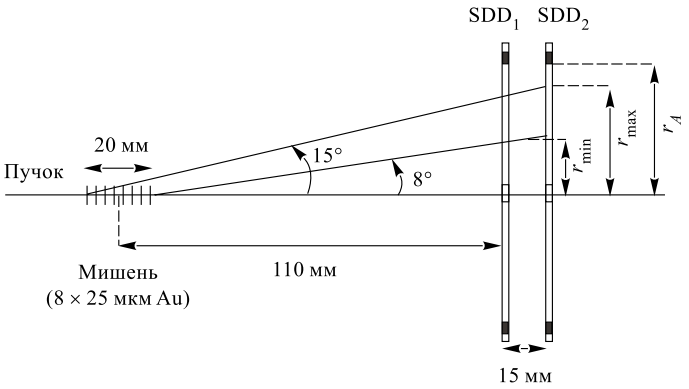



Рис. 5. Схема вершинного детектора установки CERES, включающего две силиконовые дрейфовые камеры. Отмечен угловой аксептанс детектора. Слева указана мишень, состоящая из восьми золотых дисков

весьма грубо: середина отрезка мишенной области оси $Z$. На рис. 6 показано распределение вершин, найденных по нескольким событиям на фоне дисков мишени.

Результаты прогона по 4 тыс. $\mathrm{Pb}+\mathrm{Au}$-событий позволили получить вполне удовлетворительную точность $\simeq 300$ мкм по оси $Z$, а также хорошую локальную точность трека, т. е. азимутальные и радиальные среднеква-

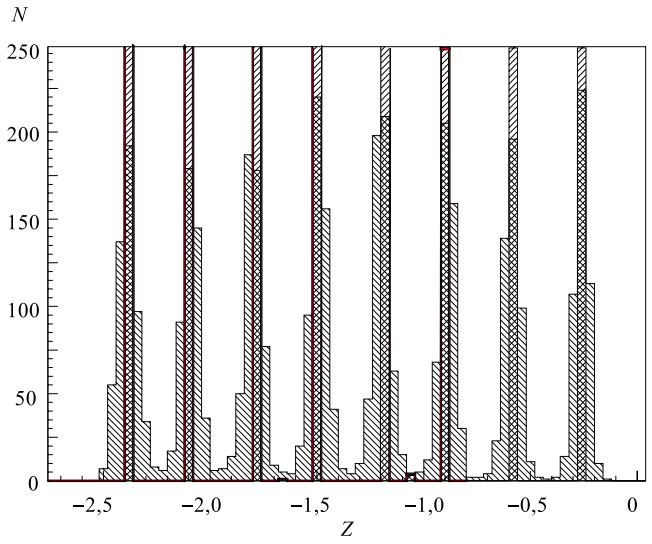



Рис. 6. Распределение вершин на фоне дисков мишени

дратичные отклонения, определенные для $\mathrm{SDD}_1$ по найденной вершине и соответствующей точке трека на $\mathrm{SDD}_2$: $\sigma_\phi \simeq 6$ мрад, $\sigma_r \simeq 100$ мкм (см. детали в [14]).

*1.3.2. Распознавание колец черенковского излучения и идентификация частиц.* Детекторы черенковского излучения типа RICH (Ring Imaging CHerenkov) широко применяются в экспериментальной ФВЭ как часть системы идентификации частиц (PID) многих экспериментов (см., например, [9–13]). Несмотря на различие конструкций RICH-детекторов, большинство из них состоит из трех основных частей: радиатора, детектора и считывателя. Черенковские фотоны, получаемые в радиаторе, фокусируются на детектирующую часть и затем эта информация считывается с помощью двумерной матрицы из многих тысяч фотоячеек. Фотоны, испускаемые детектируемой частицей, после фокусировки формируют на плоскости этой фотоматрицы кольцо (или, возможно, эллипс). Радиус получившегося кольца определяется типом (массой) частицы и ее импульсом, который обычно становится известен из других, не относящихся к RICH измерений. Таким образом, оценивая радиус, мы можем определить тип частицы — идентифицировать ее. Однако оценить радиус непосредственно путем МНК-подгонки окружности к данным измерений в таких высокоточных RICH-детекторах, как в экспериментах CERES или COMPASS [9, 15], мы не можем, поскольку эти данные, как уже указывалось выше (см. п. 1.2), являются не точками, а кластерами из смежных ячеек регистрирующей фотоматрицы (см. рис. 7, 8). В более ранней работе [16] робастный подход был применен к результатам препроцессинга с выделением хитов-центров кластеров по формуле (30). Тем не менее и в подгонке окружностей по точкам-хитам проблемы со значительным количеством шумовых хитов и фоновых хитов от близких колец могли быть успешно преодолены только при использовании робастного подхода. Для оценки радиуса $R$ и координат центра $a, b$ окружности по набору точек $(x_i, y_i)$, $i = 1, \ldots, n$, минимизировался фунционал (32) с

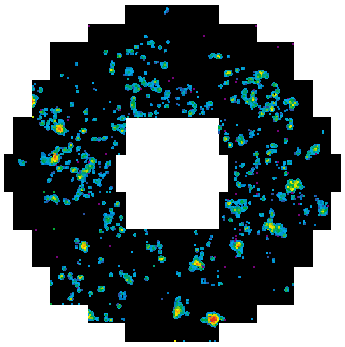


Рис. 7. Изображение данных события Au–Pb, зарегистрированных детектором CERES RICH1

$$e_i = \sqrt{(x_i - a)^2 + (y_i - b)^2} - R \tag{33}$$

и оптимальной весовой функцией (27) с $c_{\mathrm{opt}} = 0,2$. Детали линеаризации функционала можно найти в [16].

По сути, в работе [16] проведен сравнительный анализ трех робастных подходов: двух с функционалом (32), но с разными весовыми функциями —

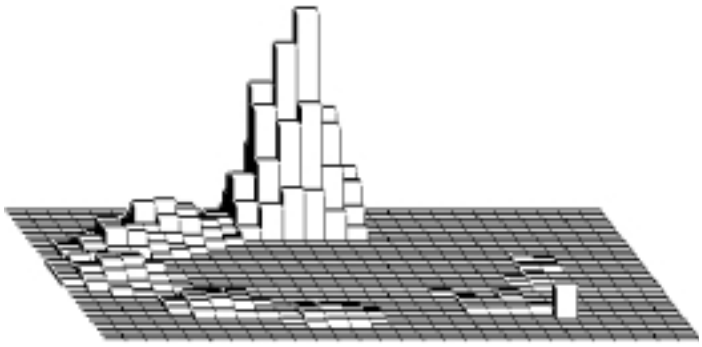


Рис. 8. Изображение смоделированных данных регистрации кольца черенковского излучения

простейшей (19) и оптимальной (27), а также с функционалом вида (16) с гауссовой функцией вклада $\rho(e)$. Последний вариант (названный SGW — summed gaussian weights) из-за сильной нелинейности функционала потребовал привлечения специального пакета программ MINUIT [17], значительно уступавшего в скорости ФГХ-процедуре даже с учетом затрат на линеаризацию функционала. Результаты на реальных Au–Pb-данных показали высокую эффективность ($\sim 95\,\%$) и примерно одинаково хорошую точность робастного метода с оптимальной весовой функцией и SGW-метода, в то время как применение простейших весов привело к потерям и в эффективности ($< 85\,\%$), и в точности.

Поскольку эксперимент CERES был специально ориентирован на поиск узких $e^+ - e^-$-пар, возникла проблема разрешения близких, перекрывающихся черенковских колец. Для одновременной подгонки двух окружностей было разработано уравнение такой комбинированной кривой четвертого порядка с шестью параметрами $a, b, c, d, R_1, R_2$:

$$F(x_i, y_i; a, b, c, d, R_1, R_2) = \\ = [(x_i - a)^2 + (y_i - b)^2 - R_1^2][(x_i - c)^2 + (y_i - d)^2 - R_2^2] = 0.$$

Оценка этих параметров потребовала минимизации нелинейного функционала

$$L(a, b, c, d, R_1, R_2) = \sum_i w_i F^2(x_i, y_i; a, b, c, d, R_1, R_2). \tag{34}$$

Его линеаризация была выполнена аналогично [16], детали вывода формул приведены в [8].

Задачи непосредственной обработки «сырых» данных черенковских детекторов, как для одной, так и для сразу нескольких окружностей, были

решены в работах [8, 18–20] с помощью кусочно-линейной аппроксимации двумерной весовой функции, зависящей также и от амплитуды сигнала. Результаты сравнительного анализа применения нескольких вариантов одномерных и двумерной весовой функций для минимизации функционала (34) на модельных данных можно найти в [20].

Совершенно иной метод был предложен для решения этой же задачи в работе [21], где применялся метод глобального случайного поиска с использованием цепи Маркова и алгоритма Митрополиса–Хастингса для ускорения поиска.

Идея весовой функции, зависящей от амплитуды, позволила предложить новый, более эффективный алгоритм идентификации частиц по радиусу черенковского угла. Для проверки гипотезы о том, какой из двух альтернативных частиц, например, $\pi$- или $K$-мезону более правдоподобно соответствует измеренный радиус, обычно на базе статистического критерия отношения правдоподобия (КОП) выбирается критическая область $r > r_0$. Известны более продвинутые методы, учитывающие ячеистую структуру данных RICH путем подсчета числа ненулевых ячеек в доверительных кольцах (ПЯДК-метод) вокруг окружностей, полученных для обеих альтернативных частиц [22]. В [20] предложен похожий подход, но с подсчетом суммы амплитуд всех ячеек в доверительной области (САДК-метод). Результаты сравнения этих методов на модельных данных для $\pi$- и $K$-мезонов приведены на рис. 9. Константа КОП для САДК-метода была выбрана так, чтобы минимизировать вероятность неправильной идентификации (вероятность ошибки

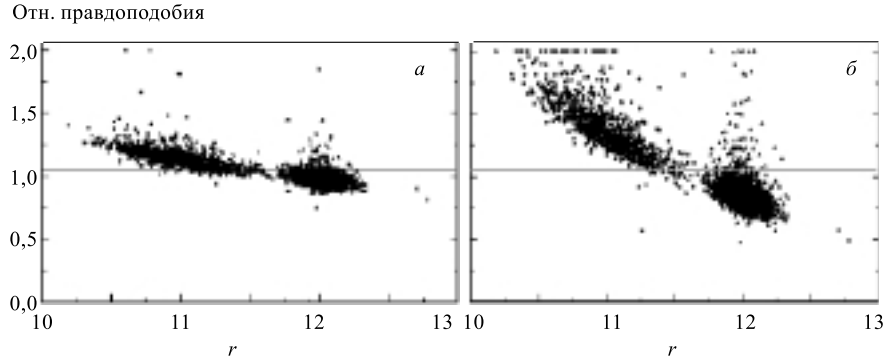



Рис. 9. Отношение правдоподобия как функция черенковского радиуса. Представлены результаты методов ПЯДК (*а*) и САДК (*б*). Верхний кластер точек на обеих частях рисунка соответствует проверяемой гипотезе. Горизонтальная линия соответствует пороговой константе критерия, так что порция верхнего кластера под этой линией представляет вероятность ошибки идентификации, в то время как часть нижнего кластера, лежащая над этой линией, представляет случаи, когда ошибочно принимается гипотеза об идентификации не той частицы (ошибка второго рода)

первого рода 1,04 %). При этом вероятность ошибки второго рода для САДК-метода оказалась равной 2,4 %, в то время как для ПЯДК-метода вероятность ошибки неправильной идентификации оказалась в три раза больше — 3,2 % при одинакой вероятности ошибки второго рода 2,3 %.

*1.3.3. Робастные оценки параметров треков.* Распознавание и оценка параметров треков элементарных частиц, получившие в последнее время короткое название *трекинг*, являются одними из основных задач обработки данных ФВЭ, т. к. именно в данных траекторных измерений содержится интересующая физиков информация об импульсных и угловых распределениях взаимодействующих частиц, необходимая для проверки основных теоретических гипотез. Выше уже отмечалось, как специфика современных экспериментов с их предельными загрузками и высокой плотностью данных на событие требует привлечения робастных методов. Отметим работы [23–25], где удачный выбор степени полинома в выражении для весовой функции типа (23) позволил успешно восстановить треки в мюонной части экспериментальной установки CMS на сложном фоне $\delta$-электронов, сопровождающих мюоны.

Укажем на еще одно специфическое проявление современных технологий детектирования, также ведущее к проблеме неоднозначности интерпретации данных, которая с успехом может быть решена робастным путем. Речь идет об измерениях, проводимых с помощью дрейфовых камер, состоящих из газонаполненных трубок, образующих плотные слои сотовой структуры. Каждая из трубок имеет центральную электродную проволоку, так что при прохождении частицы сквозь такую трубку регистрируется не только грубая координата этой проволоки, но и время дрейфа до нее облака ионов, образующихся в газе при прохождении частицы. При известной скорости дрейфа это время пересчитывается в радиус дрейфа, позволяющий вычислить координату пролета частицы до точностей порядка микрометров.

Таким образом, данные об измеренном треке в дрейфовой камере выглядят как набор малых окружностей с центрами в зарегистрированных центральных проволоках и с радиусами, равными соответствующим радиусам дрейфа. С точностью до ошибок измерения этих радиусов трек проходит по касательным ко всем этим маленьким окружностям (см. рис. 10). Главная неприятность здесь состоит в том, что знание радиуса дрейфа не позволяет определить, слева или справа от проволоки прошла частица, возникает известная *лево-право-неопределенность*. Ситуация усугубляется как из-за ложных срабатываний некоторых трубок, так и из-за их неэффективности, приводящей к пропускам некоторых измерений. В работе [26] для решения этой задачи, возникающей в системе контролируемых дрейфовых трубок (MDT — monitored drift tubes) в магнитном поле установки ATLAS [27], применен гибридный алгоритм, сочетающий робастный метод, позволяющий быстро обработать $\sim 90\,\%$ событий, и более медленный метод эластичной нейронной сети (см. разд. 3) для оставшихся $\sim 10\,\%$ событий.

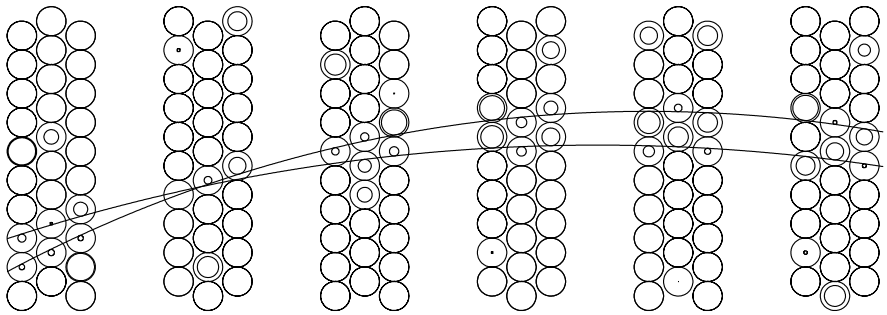


Рис. 10. Пример типичного события в дрейфовой камере в магнитном поле

Основные проблемы при применении робастного метода для данных дрейфовых камер:

• распознавание каждого из треков, т. е. выделение в отдельный массив набора измерений (тройка чисел на каждую измеренную точку — две координаты центральной проволоки и радиус дрейфа), относящихся к данному распознанному треку из множества треков, составляющих анализируемое событие, причем число их заранее неизвестно;

• определение начальных значений параметров, описывающих трек;

• подгонка треков с учетом лево-право-неопределенности.

Ограничимся для наглядности рассмотрением одной проекции $X0Z$.

Распознавание треков достаточно провести по множеству измеренных центров проволок $(x_i, z_i)$, $i = \overline{1,n}$. Распознавание и одновременная грубая оценка начальных значений параметров треков выполнялись с помощью последовательного варианта *дискретного преобразования Радона–Хафа* (см., например, [28] или описание быстрой версии преобразования для прямых треков [29]). Это преобразование переводит линии в пространстве измерений в точки в пространстве параметров.

Для данных дрейфовых камер в однородном магнитном поле треки описываются дугами окружностей. Поэтому перебираются все *допустимые* неколлинеарные триплеты $(x_i, z_i)$, $i = 1, 2, 3$, по которым вычисляются координаты центра проводимой через них окружности

$$\begin{aligned} x_c &= \frac{1}{2}\frac{(x_2^2 - x_3^2 + z_2^2 - z_3^2)(z_1 - z_2) - (x_1^2 - x_2^2 + z_1^2 - z_2^2)(z_2 - z_3)}{(x_2 - x_3)(z_1 - z_2) - (x_1 - x_2)(z_2 - z_3)}, \\ z_c &= \frac{1}{2}\frac{(x_1^2 - x_2^2 + z_1^2 - z_2^2)(x_2 - x_3) - (x_2^2 - x_3^2 + z_2^2 - z_3^2)(x_1 - x_2)}{(x_2 - x_3)(z_1 - z_2) - (x_1 - x_2)(z_2 - z_3)} \end{aligned} \tag{35}$$

и ее радиус

$$R = \sqrt{(x_i - x_c)^2 + (z_i - z_c)^2},\ i = 1, 3.$$

Допустимость понимается в том смысле, что принимаются только те триплеты, у которых $R_{\min} < R < R_{\max}$, а центры также лежат в заданной области. 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. Из-за наличия ошибок измерений этот пик несколько размазывается, что сказывается на размерах биннинга (разбиения на ячейки) при гистограммировании получаемых параметров.

Гистограммирование разумнее выполнить в два этапа: сначала строится $2D$-гистограмма центров и находятся все пики в ней, превышающие заданный порог (см. рис. 11). Затем для каждого найденного центра $(x_c, z_c)$ выбираются точки внутри полосы $R_{\min} < \rho_i < R_{\max}$, $\rho_i = \sqrt{(x_i - x_c)^2 + (z_i - z_c)^2}$, и также гистограммируются. Максимальный из пиков гистограммы принимается за оценку радиуса. Процедуры выбора порогов и тестирования полученных параметров достаточно сложны. Некоторые рекомендации можно найти в [26, 97].

Получив таким образом приблизительные начальные значения параметров треков, мы можем начать их последовательную робастную подгонку с

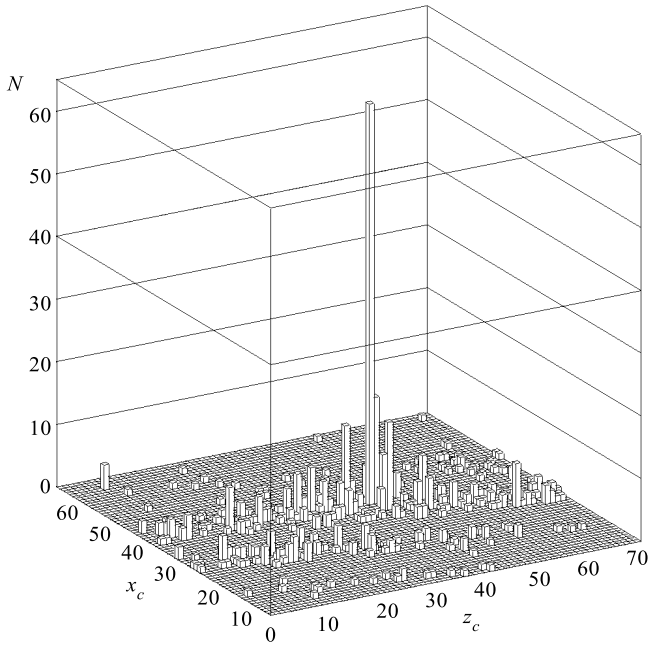



Рис. 11. $2D$-гистограмма центров для одного трека из 10 точек со 100 % зашумлением

учетом лево-право-неопределенности. С этой целью для каждого из центров проволок $(x_i, z_i)$, отобранных при распознавании в массив точек очередного трека, по соответствующему радиусу дрейфа $r_i$ вычислялись две возможные точки прохождения трека вблизи этой проволоки $(x_{1i}, z_{1i})$ и $(x_{2i}, z_{1i})$. При подгонке по такому удвоенному множеству точек весовые функции робастного метода вида (23) при правильно выбранном параметре $c_T$, учитывающем как диаметр трубки, так и ошибку измерений $r_i$, позволяют автоматически определить значимость этих двух альтернативных точек и, соответственно, выполнить подгонку по наилучшему из вариантов.

*1.3.4. Проблемы алайнмента.* Любой детектор состоит из множества отдельных составляющих его частей — поддетекторов, модулей, положение которых при сборке всего детектора неизбежно нарушается из-за случайных смещений от их идеальных позиций, предусмотренных проектом. И хотя проект обычно гарантирует очень высокую точность измерений в каждом отдельном модуле, эти смещения могут привести к значительным потерям в точности. Таким образом, возникает задача алайнмента, решение которой должно обеспечить следующее:

• используя данные обычных измерений, определить возможные пространственные дисторсии, т. е. смещения и повороты элементов детектора, каждый из которых рассматривается уже как некое твердое тело;

• выполнить математическую коррекцию найденных дисторсий с помощью специального корректирующего преобразования.

В обзоре [30] описаны похожие задачи для оптических трековых детекторов, называемые задачами калибровки. Их цель — использовать измерения особого эталонного изображения для преобразования измеряемых детектором величин из внутренней системы координат детектора в унифицированную лабораторную систему координат.

В принципе, по самой идее — использовать данные измерений для определения и математической коррекции дисторсий прибора — алайнмент мог бы рассматриваться как некая разновидность калибровки, однако физики предпочитают различать эти понятия как из-за высокой иерархичности задач алайнмента, так и из-за того, что в больших электронных детекторах вместо эталона приходится пользоваться данными обычных измерений. Поскольку весь детектор может представлять собой огромную иерархическую систему из большого числа сложных компонентов, процедура алайнмента также неизбежно распадается на локальные процедуры алайнмента для отдельных поддетекторов и на глобальный алайнмент всего детектора в целом. Подход, называемый DDA (data driven approach), когда вычисления управляются данными, также характерен для задач алайнмента, хотя он по необходимости используется теперь и для калибровки некоторых частей детекторов (как, например, при получении калибровочного преобразования времен дрейфа в системе дрейфовых камер в радиусы дрейфа [31]).

В силу многообразия как экспериментов (коллайдерных или с фиксированной мишенью), так и структур соответствующих детекторов применяются разные методы локального и глобального алайнмента (см., например, [32–37]).

Здесь мы ограничимся кратким описанием схемы локального алайнмента вершинного силиконового детектора SVT установки STAR, где использование робастности сыграло важную роль в достижении высокой эффективности предлагаемого метода.

Трехслойная конфигурация SVT включает в себя 216 кремниевых дрейфовых камер, собранных в три концентрических слоя вокруг пучка внутри времяпроекционной камеры, так что на внутреннем, среднем и внешнем слоях находятся 32, 72 и 112 пластин соответственно. Рассматривая каждую из кремниевых дрейфовых камер как твердое тело, ее общее смещение от идеального положения можно представить шестью параметрами (три смещения и три поворота), которые распадаются на четыре различные категории, различным образом влияющие на правильность определения положения взаимодействия:

- смещения в плоскости камеры (два параметра),
- вращение в плоскости камеры (один параметр),
- изменение радиального расстояния (один параметр),
- вращения не в плоскости камеры (два параметра).

Хотя каждая из детекторных пластин имеет 6 степеней свободы, что, соответственно, дает 18 параметров, однако, как показано в [36], реально можно восстановить только 9 из них. Для определения оставшихся параметров требуется дополнительная информация, получаемая в силу того, что треки, по которым будет вестись алайнмент, выходят из одной вершины и проходят одновременно через несколько камер. Следует учитывать, что в отсутствие магнитного поля треки прямолинейны, но их измерения содержат погрешности, вдобавок требуется еще учесть и наличие многократного рассеяния.

Кратко процедура алайнмента может быть описана следующим образом [38]. Пусть измерены точки $(x_d,\ x_a)$, где $x_d$ и $x_a$ — измерение положений соответственно в дрейфовом и анодном направлении (т. е. в плоскости самой детектирующей пластины). Используя эту информацию на «входе», вначале необходимо вычислить трехмерные координаты точки пересечения $(P)$ в принятой глобальной системе координат STAR $(x,\ y,\ z)$, используя измеренные (наблюдаемые) положения и ориентации SVT-детектора. После чего задействуется процедура восстановления прямолинейных треков и формирования триплетов точек $(P_1,\ P_2,\ P_3)$. По приблизительно прямым трекам, восстановленным таким образом, определяется положение главной вершины события $(V_x,\ V_y,\ V_z)$. После чего трек и вершина фитируются вместе для увеличения точности определения параметров траектории (т. е. для уменьше-

ния среднего расстояния между подогнанной траекторией и точками трека). Таким образом, в качестве основного показателя качества алайнмента была выбрана величина $\chi^2$ треков. Для случая идеальной выстроенности всех детекторных пластин остатки должны быть в среднем равны нулю, а величина трекового $\chi^2$ должна отражать только эффекты ограниченного позиционного разрешения самого детектора и многократного рассеяния. Геометрические параметры пластин варьировались при помощи быстрой версии метода градиентного спуска [38] для определения набора параметров, минимизирующих трековые остатки. Ограничивающее уравнение, создаваемое наличием общей вершины, включено через добавление точки первичного взаимодействия к каждому треку. Итерационная процедура градиентного спуска продолжается до тех пор, пока величина трекового $\chi^2$ продолжает уменьшаться, прерываясь, когда дальнейшее варьирование параметров не приводит к существенному улучшению треков. С этого момента параметры пластин (позиции и ориентации) считаются вычисленными и сохраняются в виде, требуемом для работы другого математического обеспечения. Процедура алайнмента на этом считается завершенной.

В разработке этой процедуры наиболее сложно было составить минимизируемый функционал, который должен был зависеть от 9 параметров каждой из 216 кремниевых дрейфовых камер и учитывать ограничения, связанные с тем, что все треки события выходят из одной вершины. Кроме того, следовало учитывать, что дисторсии отдельных элементов SVT, а также статистические погрешности измерений записывались в локальной системе координат этого элемента, но в функционал они должны были входить, будучи переведенными в общую для детектора систему координат. С учетом большой множественности событий в эксперименте STAR и упоминавшегося наличия погрешностей измерений и фоновых отсчетов использовался робастный функционал с весовой функцией Тьюки (23).

Как видно из табл. 1 [38], робастный подход дает значительно лучшее разрешение по каждой координате.

Детали многопараметрической минимизации сильно нелинейного функционала можно найти в [38, 39, 100].

*Таблица* 1. **Точность восстановления вершины**

| Метод | $\sigma_x$, мкм | $\sigma_y$, мкм | $\sigma_z$, мкм |
|---|---|---|---|
| Простой | 23,43 | 10,35 | 10,91 |
| Робастный | 7,15 | 4,92 | 5,36 |

## 2. ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ

В последние десятилетия резко возросло число исследований по применению искусственных нейронных сетей (ИНС) в самых различных областях — технике, геологии, медицине, бизнесе и особенно в экспериментальной физике, где ИНС широко применяются для анализа данных. Секции нейронных сетей появляются в программах практически всех конференций, посвященных применению искусственного интеллекта или вычислительных машин в физике, растет число публикаций о приложениях нейронных сетей в таких ведущих физических журналах, как «Nuclear Instruments and Methods» или «Computer Physics Communications». Со времени выхода обзора по применению ИНС в экспериментальной физике [40] только сотрудниками ОИЯИ выпущено свыше сорока публикаций по этой тематике. Наиболее интересные из них рассмотрены в настоящем разделе.

Напомним вкратце об основных понятиях ИНС (см., например, [40]):

• нейроны представляют собой простые логические устройства, характеризуемые

— уровнем активации;

— топологией связей нейронов друг с другом;

— мерой взаимодействия с другими нейронами, называемой синаптической силой связи или весом. Веса этих связей различны и могут определяться в зависимости от решаемой задачи;

— выходным уровнем, который связан с уровнем активации посредством некоторой функции обычно сигмоидального типа;

• вся система состоит из очень большого числа одинаковых нейронов, причем результат работы ИНС мало чувствителен к характеристикам конкретного нейрона;

• система допускает возможность параллельной обработки информации. Таким образом, если обозначить сигнал, исходящий от $k$-го нейрона сети, как $x_k$, вес синаптической связи $j$-го и $k$-го нейронов — как $w_{jk}$, то общий входной сигнал, поступающий на $j$-й нейрон, равен

$$h_j = \sum_k w_{jk} x_k. \tag{36}$$

Выходной сигнал $j$-го нейрона получается путем воздействия на этот суммарный сигнал активационной функцией

$$y_j = g(h_j), \tag{37}$$

где $g(u)$ — либо пороговая функция, либо сжимающая функция сигмоидального вида

$$g(u) = \frac{1}{1 + \mathrm{e}^{-u\lambda}}, \tag{38}$$

как на рис. 12.

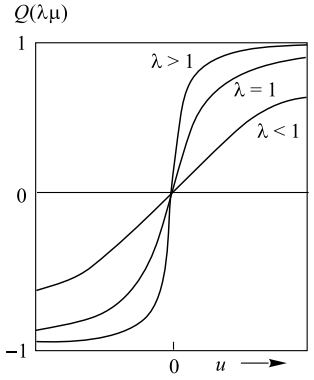



Рис. 12. Общий вид активационной функции в зависимости от параметра $\lambda$

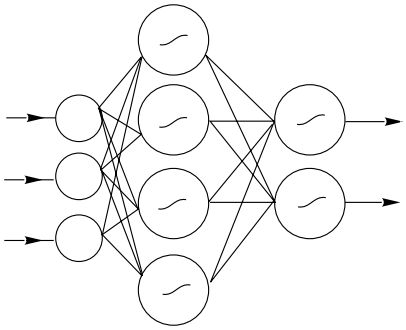


Рис. 13. Схема трехслойного персептрона

Ключевыми характеристиками сети являются *тип связей* между нейронами и *динамика эволюции сети*, определяемая функцией активации нейронов и правилом изменения весов в процессе этой эволюции.

ИНС, наиболее распространенные в физике, определяются двумя типами связей: прямоточные сети без обратных связей, например, многослойные персептроны (МСП) (см. рис.13), и полносвязные сети, в которых все нейроны связаны друг с другом, как в нейронной сети Хопфилда (ХНС).

Нейронные сети успешно применяются везде, где нужно решать задачи классификации, прогнозирования или распознавания. Этот успех определяется следующими возможностями.

• Широкий спектр решаемых задач. Применение ИНС носит поистине междисциплинарный характер, т. к. позволяет успешно решать:

— многие *нелинейные* задачи, в то время как классические вычислительные методы, как правило, предполагают линейность задач, или, если и допускают нелинейность, то с ограничениями по величине, чтобы гарантировать сходимость различных итерационных процедур;

— задачи с данными стохастической природы. Робастность ИНС проявляется благодаря возможности их обучения на большом числе тренировочных выборок.

• Сравнительная простота структуры ИНС и простота применения. Это особенно касается физики, где обычно квалификации дипломированного физика-экспериментатора вполне хватает для использования программной или схемной реализации ИНС, и не возникает проблем с моделированием необходимой последовательности данных, требуемых для обучения сети.

Простота структуры ИНС типа многослойного персептрона стимулировала многих исследователей на создание универсальных пакетов для генерации МСП по заданному числу слоев и нейронов в них, позволяющих также реализовать один из методов обучения сети. Примечательно, что один из первых таких нейропакетов программ JETNET был разработан именно физиками из Лундского университета [41]*.

По этим же причинам нескольким электронным фирмам уже в начале 90-х гг. удалось осуществить аппаратные реализации ИНС широкого применения в виде интегральных чипов, работающих параллельно и допускающих настройку сети на заранее смоделированную конфигурацию (см., например, [85] или обзор [42]). Значительное количество статей по применению различных нейрочипов (ETANN, TOTEM, ZISC036) и даже специального алгоритмического языка CNAPS для управления ими можно найти в специальном выпуске журнала «Nuclear Instruments & Methods in Physics Research» [43].

**2.1. Применение многослойных персептронов.** Рассмотрим, как можно обучить МСП для классификации или распознавания. Входному вектору $\mathbf{x} = (x_1, x_2, \ldots, x_n) = \{x_{inp}\}$ МСП ставит в соответствие выходной вектор $\mathbf{y} = \{y_j\}$. Например, для трехслойного персептрона входные сигналы сначала преобразуются в нейронах скрытого слоя как

$$h_k = g(H_k), \;\; H_k = \sum_i w_{ik} x_i; \tag{39}$$

а потом сигналы от нейронов скрытого слоя преобразуются нейронами выходного слоя как

$$y_j = g(H_j), \;\; H_j = \sum_k w_{kj} h_k. \tag{40}$$

Это преобразование $\mathbf{x} \Rightarrow \mathbf{y}$ полностью описывается с помощью синаптических весов $\{w_{ik}\}$, $\{w_{kj}\}$, которые и должны быть найдены, чтобы использовать МСП для решения какой-то конкретной задачи. Веса можно определить, если имеется набор данных с уже известными свойствами, так называемая *обучающая выборка*, состоящая из пар векторов $(\{x_i\}^{(m)}, \; \{z_j\}^{(m)})$, $m = \overline{1, M}$, где $M$ — объем такой выборки. Обучение МСП наиболее часто применяемым методом обратного распространения ошибок (ОРО) основано на сравнении векторов этих пар, т. е. известного результата классификации

*Одной из наиболее рекламируемых в настоящее время доступных в России программ генерации ИНС с расширенными средствами препроцессинга данных считается коммерческий пакет «Statistica Neural Network», поставляемый фирмой «StatSoft. Inc» (см. http://www.statsoft.ru).

$\mathbf{z}^{(m)}$ и выхода МСП $\mathbf{y}^{(m)}$ с помощью квадратичного функционала

$$E = \sum_m \sum_j (y_j^{(m)} - z_j^{(m)})^2 \Rightarrow \min, \tag{41}$$

с использованием весов $\{w_{ik}\}$, $\{w_{kj}\}$ как минимизационных параметров. Решение обычно выполняется методом наискорейшего спуска. Приравнивая к нулю производные (41) по весовым параметрам, получаем итерационные правила изменения весов на каждой эпохе обучения $t$:

$$\Delta w_{kj}^{(t)} = -\eta (y_i^{(n)} - z_i^{(t-1)}) g'(y_i^{(t-1)}) h_k^{(m)} \tag{42}$$

для весов выходного слоя и

$$\Delta w_{ik}^{(t)} = -\eta \sum_i w_{kj} g'(y_j^{(t-1)}) g'(h_k^{(t-1)}) x_k^{(t-1)} \tag{43}$$

для весов скрытого слоя. Здесь $\eta$ — параметр скорости обучения [44].

Следует заметить, что проблема ускорения обучения МСП является весьма важной в практических приложениях. Среди многих известных способов ее решения отметим следующие:

• использование вычислительных методов ускорения минимизации целевого функционала сети;

• преобразование исходных данных с целью повышения их информативности;

• оптимизация структуры МСП для сокращения числа нейронов на всех слоях (главным образом, в скрытом слое) и, соответственно, числа связей между ними, т. е. числа весов, по которым ведется минимизация;

• смены самого способа разбиения пространства признаков при классификации, например, путем перехода к так называемым RBF-сетям (RBF — radial basis function).

Рассмотрим эти методы подробнее.

*2.1.1. Методы минимизации целевого функционала ИНС, отличные от градиентного.* Поскольку в зависимости от исходных данных поверхность функционала (41) может иметь нерегулярности типа оврагов, предлагалось несколько способов минимизации и в этих ситуациях: метод сопряженных градиентов, метод Левенберга–Маркара [45,46] или добавление к правым частям уравнений (42), (43) регуляризационных $\Delta$-коррекций весов вида $\alpha \Delta w_{ik}^{(m)}$ и $\alpha \Delta w_{kj}^{(m)}$ с $\alpha < 1$.

Метод ОРО с подобными $\Delta$-коррекциями применялся в работе [47] для отбора событий с $B$-мезонами в калориметрическом триггере, а также в [48] для идентификации полезных событий в эксперименте DISTO.

В работе [49] было проведено сравнительное исследование качества обучения с помощью обычного ОРО-метода наискорейшего спуска, реализованного в программе JETNET [41], и подхода, основанного на использовании метода ньютоновского типа, при котором весовые поправки вычисляются как

$$\Delta w = -\eta(\bigtriangledown^2 E)^{-1} \bigtriangledown E, \tag{44}$$

где гессиан $\bigtriangledown^2 E$ вычисляется из

$$\frac{\partial^2 E}{\partial w_{ik} \partial w_{kj}} = y_i \frac{\partial h_k}{\partial w_{ik}} + h_k \frac{\partial y_j}{\partial w_{kj}}.$$

В чистом виде этот метод неприменим для данной задачи, поскольку гессиан может вырождаться вблизи минимума, что ведет к некорректности задачи. Попытки применить регуляризацию с малым параметром $\alpha$ [51] не привели к результату лучшему, чем в JETNET. Задачу удалось решить путем замены формулы (44) на

$$\Delta w = -\eta H^{-1} \bigtriangledown E,$$

где матрица $H$ невырождена и положительно определена: $H = \bigtriangledown^2 E + \mu I$ с таким $\mu$, чтобы гарантировать, что собственное значение $H \geq \delta > 0$. Выбор $\delta = 1$ дал лучший, чем для ОРО, результат при обучении нейросети, хотя и потребовал больших затрат компьютерного времени.

В дальнейшем в задаче распознавания $B$-мезонов по данным вершинного детектора в эксперименте МЧС этот подход позволил предложить алгоритм, сокращающий фоновые процессы примерно в 40 раз [52].

*2.1.2. Преобразование исходных данных для повышения их информативности.* Удачный выбор переменных, подаваемых на вход МСП, часто является определяющим моментом для успешного решения задачи. Несмотря на то, что у физиков, как правило, есть замечательные развитые программные средства генерации событий, позволяющие смоделировать любое их число с учетом реальных условий экспериментов, т. е. обеспечить репрезентативность и сбалансированность обучающей выборки, очень часто удается преобразовать этот набор исходных данных так, чтобы повысить их информативность и по возможности сократить их размерность.

В работах [53, 88] приведен наглядный пример, когда простейшее преобразование — упорядочивание выборочных значений — позволяет значительно ускорить обучение МСП (эффект «мгновенного обучения») в достаточно сложной задаче классификации данных, выбранных из двух нормальных распределений с одинаковыми средними и различными дисперсиями.

Интересный пример дан в [89], где для разделения протонов и пионов при космических энергиях выше 100 ГэВ по данным их ионизационных потерь, зарегистрированных в шести детекторах, сначала вычислялись значения $\omega_6^5$-критерия [54], которые затем подавались на вход МСП, обученного

на $10^4$ смоделированных событиях. Эффективность правильной классификации на модельных данных составляет 93,5 % при ошибке второго рода $\sim 7\,\%$.

Наиболее эффективным способом сокращения размерности обучающей выборки без потери ее информативности является *метод главных компонентов* (МГК) [55]. Ниже в п. 2.1.4 приводится пример с применением МГК к предварительной обработке оцифрованного изображения для последующего ввода в ИНС. Дело в том, что любая попытка использования ИНС для анализа объектов большой размерности* наталкивается на необходимость работы с сетями из десятков тысяч нейронов. При этом содержательная часть анализируемого объекта занимает в пространстве входных признаков незначительное подпространство много меньшей размерности.

МГК дает нам путь проектирования векторов $\mathbf{x}$ данных на такое подпространство при сохранении наиболее адекватных признаков анализируемого объекта, используя информацию об их взаимной корреляции, т. е. ковариационную матрицу $\Sigma_{\mathbf{x}} = \operatorname{cov}(x_i x_k)$.

Для этого применяется ортогональное преобразование Карунен–Лоэва $L = \{l_{ki}\}$ (см., например, [56], с. 220), которое преобразует $\Sigma_{\mathbf{x}}$ к диагональной форме

$$\Sigma_{\mathbf{y}} = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{pmatrix}, \tag{45}$$

где собственные значения $\lambda_i$ пронумерованы по убыванию. Таким образом, можно оставить только $m$ из них: $\lambda_1, \lambda_2, \cdots, \lambda_m\,(m \ll p)$ наиболее значимых, и мы теперь можем выразить исходные данные $\mathbf{x}$ через эти главные компоненты:

$$x_i \cong l_{1i} y_1 + l_{2i} y_2 + \ldots + l_{mi} y_m. \tag{46}$$

МГК успешно применялся для предварительного сжатия трековой информации на спектрометре R603 еще в 1973 г. [57] (см. также более поздние работы [58]). Пример применения МГК для предобработки входных данных ИНС приведен в п. 2.1.4.

*2.1.3. Оптимизация структуры МСП.* Поскольку минимизация функционала (41) ведется по весам связей между нейронами МСП, число которых определяется как произведение чисел нейронов в каждом слое, то, очевидно, чем компактнее сеть по числу нейронов в ней, тем легче ее обучить. Более

*Заметим, что даже оцифрованное черно-белое изображение на скромном растре $100 \times 100$ уже дает входной вектор с $10^4$ компонентами.

того, чтобы оценить весовые параметры с достаточной точностью, длина обучающей последовательности должна быть как минимум на порядок больше числа этих параметров. Насущность этой проблемы часто подчеркивается термином «проклятие размерности» ИНС.

Выше мы уже обсудили, как минимизировать число входных признаков, определяющих число $n_1$ нейронов первого слоя. Если, например, МСП осуществляет классификацию на $m$ классах, то, казалось бы, число нейронов выходного слоя должно определяться как $n_3 = \log_2 m$. Однако результаты работы сети, организованной таким образом, как говорят, «под завязку», не очень надежны. Для повышения достоверности классификации желательно ввести избыточность путем выделения каждому классу одного или даже нескольких нейронов в выходном слое, каждый из которых обучается определять принадлежность образа к классу со своей степенью достоверности, например, высокой, средней и низкой. Такие МСП позволяют проводить классификацию входных образов, объединенных в нечеткие (размытые или пересекающиеся) множества. Это свойство приближает подобные МСП к условиям реальной жизни. Существуют также эвристические рекомендации по поводу выбора числа нейронов в скрытом слое типа $n_2 = 2n_1 + 10$ [50] или $n_2 = (n_1 + n_3)/2$. Следует, однако, заметить, что избыточность в скрытом слое (слоях) вредна, т. к. она увеличивает без необходимости число параметров, по которым ведется минимизация. Для устранения избыточных нейронов применяется процедура «обрезания» (prunning), т. е. устранения тех нейронов скрытого слоя, у которых веса связей с выходными нейронами малы (меньше выбранного порога).

Обычно отсутствие более научно обоснованных рекомендаций по выбору структуры многослойных ИНС не очень волнует физиков, в известной степени избалованных имеющимися у них возможностями использовать программы типа JETNET для генерации сетей любой конфигурации, а также тем, что благодаря программам моделирования событий, таким как GEANT-3.0 или HIJING, практически не возникает ограничений на длину обучающей последовательности, которая фактически определяет точность вычисления весов МСП. Однако в задачах экспериментальной биологии и особенно медицины, где достоверная модель изучаемых зависимостей может и не существовать, приходится вести обучение и проверять качество работы обученной сети часто на небольшом экспериментальном материале. В этих условия выбор оптимальной структуры сети может играть ключевую роль.

В качестве примера возьмем работу [59], в которой описана экспертная система для диагностики пневмонии и прогнозирования ее исхода. Естественно, что целью разработки была не попытка замены лечащего врача, а создание системы для помощи ему в принятии решения.

В системе использовались две независимые сети различной конфигурации. Это было необходимо для того, чтобы прогноз не зависел от диагноза.

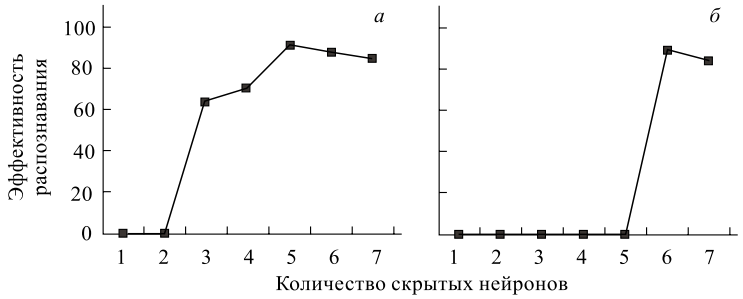



Рис. 14. Эффективность распознавания как функция числа скрытых нейронов при двух (*а*) и трех (*б*) выходных нейронах

С целью разработки системы, удобной для практического врача, было обследовано 300 пациентов с внебольничной пневмонией и 50 практически здоровых лиц. При диагностике пневмонии врачи используют 35 признаков болезни. С целью сокращения их числа был проведен предварительный анализ многих клинико-лабораторных, рентгенологических показателей, что позволило выделить 9 прогностически значимых признаков: возраст, температура, количество лейкоцитов в периферической крови, количество сегментов легкого с воспалительной инфильтрацией и др. Причем первоначально в качестве одного из входных параметров использовался стаж курения и не учитывался алкоголизм. Однако выяснилось, что сеть стала путать тяжелобольных и легкобольных. После того как добавили к сопутствующим патологиям алкоголизм, количество ошибок уменьшилось. Оказывается, хронический алкоголизм вкупе с хроническим бронхитом в большинстве случаев оказывает то самое утяжеляющее воздействие на течение пневмонии, которое и отличает тяжелобольного от легкобольного.

Были исследованы 20 различных конфигураций сетей с целью найти оптимальную, обеспечивающую наилучшую эффективность распознавания. Результаты показаны на рис. 14. Нулевая эффективность означает, что процедура обучения сети не сходится. Данная ситуация возникает при недостаточном количестве скрытых нейронов. При избыточном количестве скрытых нейронов сеть будет просто запоминать факты, а не обобщать их, что также приводит к уменьшению точности распознавания. На основании исследования была выбрана следующая оптимальная конфигурация диагностической сети: 9 входных, 6 скрытых и 3 выходных нейрона. Для прогнозирующей сети оптимальной явилась конфигурация: 9 входных, 6 скрытых, 1 выходной нейрон. Как показала проверка на контрольной группе, эффективность работы экспертной системы при выбранных структурах сетей составила 95,2 %.

*2.1.4. RBF-сети.* Нетрудно показать, что схема классификации на два класса с помощью простого МСП с пороговой функцией активации имеет

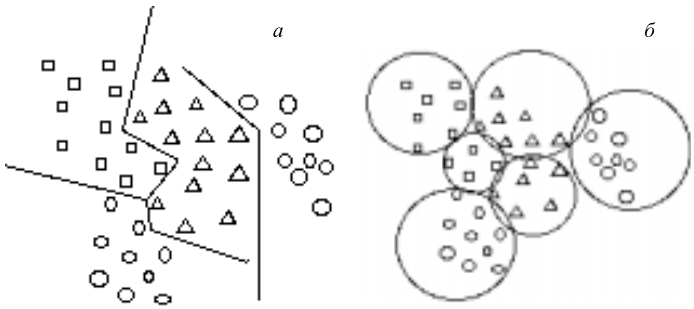



Рис. 15. Примеры разделения трех сильно пересекающихся классов с помощью: *а*) обычного многослойного персептрона; *б*) RBF-сети

простую геометрическую интерпретацию: в пространстве $H$ задана гиперплоскость $\sum_j w_{ij}^{(2)} h_j = 0$, которая делит пространство на два полупространства. Считается, что если вектор находится по одну сторону от поверхности (это значит, что для него $y_j$ из (40) выше порога), то он относится к первому классу, если же по другую — ко второму. Таким образом, то, как персептрон формирует понятия (или проводит классификацию), является процессом нахождения набора гиперплоскостей, которые обеспечивают ограничение или оконтуривание каждого из множеств, представляющих собой отдельные классы.

Как видно из (36), нейрон осуществляет преобразование вектора в действительное число путем скалярного умножения подаваемого на вход вектора $\mathbf{x}$ на вектор весовых коэффициентов $\mathbf{w}_{\cdot k}$. В алгоритме обучения именно скалярное произведение векторов играет большую роль, т. к., задавая совокупность направлений в пространстве, оно задает набор гиперплоскостей, обеспечивающих простое и интуитивно понятное разделение на классы.

Однако ограничение некоторого подмножества набором гиперплоскостей не является единственным способом разделения множеств. Аналогичного и не менее естественного результата можно достигнуть, описав каждое из множеств при помощи полного покрытия некоторым набором гиперсфер (см. рис. 15). Нейронные сети, реализующие такую классификацию, называются RBF-сетями.

В RBF-сетях вместо скалярного произведения двух векторов вычисляется расстояние между ними, т. е. в пространстве этих векторов вводится та или иная метрика $\rho(\mathbf{H}, \mathbf{w}_{\cdot k})$. В зависимости от задачи в качестве метрики можно использовать сумму модулей разностей компонентов $\sum_k |h_k - w_{ik}|$ (так называемое манхэттенское расстояние), или максимум этих модулей $\max_i |h_k - w_{ik}|$, или так называемое расстояние Махаланобиса $d^2(\mathbf{H}, \mathbf{w}_{\cdot k}) =$

$(\mathbf{H} - \mathbf{w}_{\cdot k})^{-T}\Sigma^{-1}(\mathbf{H} - \mathbf{w}_{\cdot k})$, где $\Sigma$ — оценка функции ковариации этих векторов. В работах [60, 61] предлагается квадратичная метрика

$$L_2 = \sum_k (h_k - w_{kj})^2. \tag{47}$$

Вместо функции активации (38) в таких сетях используется пороговая функция (как в [60, 61]) или чаще гауссиан.

RBF-сеть имеет промежуточный слой из радиальных элементов, каждый из которых воспроизводит гауссову поверхность отклика. Как показано в [62], для моделирования произвольной функции необходим лишь один промежуточный слой, содержащий достаточное число радиальных элементов. На выход сети подается их линейная комбинация (т. е. взвешенная сумма гауссовых функций) с линейными функциями активации. Такие RBF-сети имеют ряд преимуществ перед сетями МСП. Во-первых, как уже сказано, они моделируют произвольную нелинейную функцию с помощью всего одного промежуточного слоя и тем самым избавляют нас от необходимости решать вопрос о числе слоев. Во-вторых, благодаря линейности вычислений с параметрами в выходном слое минимизация идет быстро и без трудностей с локальными минимумами, так мешающими при обучении МСП. Поэтому сеть RBF обучается очень быстро (на порядок быстрее МСП). Нейрочипы, реализующие RBF-сети, уже разработаны (ZISC036) и применяются в триггерных системах экспериментальной ФВЭ [85].

Тем не менее МСП все еще более популярны, особенно в физических приложениях, чем RBF-сети, в силу ряда причин. По-видимому, последние менее исследованы. Считается, что когда число входов велико, «проклятие размерности» довлеет над ними в еще большей степени, требуя значительного препроцессинга для оценки числа классов и центров радиальных нейронов (см., например, [63, 85]). В этой связи стоит указать на удачный опыт использования специально организованной RBF-сети с пороговой активационной функцией, примененной в задаче распознавания оцифрованных изображений [60, 61].

Эта проблема важна для многих приложений, например:

- обработка фотографий из космоса,
- биология: анализ хромосом,
- медицина: кардиология, пульмонология,
- физика высоких энергий: распознавание треков заряженных частиц, колец черенковских излучений,
- автоматическая обработка рукописных текстов,
- распознавание изображений человеческих лиц.

Последняя задача была выбрана для исследования, поскольку в ней воплощались проблемы, типичные для остальных приложений, она была достаточно хорошо изучена [64–67], кроме того, существовала доступная через

Интернет достаточно представительная база данных с набором из 400 изображений лиц [68], сфотографированных в условиях, сходных с теми, что входили в технические требования задачи. Эти требования включали разработку алгоритма для быстрого и надежного распознавания фронтальных нецветных изображений, полученных видеокамерой с 8-разрядной градацией по яркости, пригодного для реализации на недорогой персональной ЭВМ. При этом требовалось обеспечить распознавание лиц в реальных условиях, когда неизбежны небольшие вариации поз, изменения в выражениях лиц, прическах, небритости у мужчин и т. д. Сложность задачи обуславливалась как технической проблемой преобразования непрерывного изображения в растр порядка $80 \times 100$ пикселей, порождающий 8000 входных нейронов сети, так и программными трудностями организации обучения RBF-ИНС с таким немыслимым для обычного многослойного персептрона числом нейронов. Потребовалось также оптимизировать работу сети, чтобы с добавлением нового лица в обучающее множество она быстро «доучивалась», а не переучивалась снова.

В соответствии с этими требованиями была предложена RBF-сеть с пороговой активационной функцией, с метрикой $L_2$ и двумя скрытыми слоями нейронов. В этой ИНС осуществлялась *двухуровневая схема распознавания образов*. Первый из скрытых слоев был предназначен для выполнения предварительного сокращения числа входных признаков путем их грубого усреднения по классам с помощью реализации обычной линейной МСП-схемы (36), где в качестве весов использовались некоторые постоянные, заранее вычисленные коэффициенты.

Наиболее существенным является отображение, которое осуществляется радиальными нейронами третьего слоя сети, т. к. оно должно давать ответ на вопрос о принадлежности вектора, или, что то же самое, точки гиперпространства, области, описываемой гиперсферой, и, таким образом, сводится к пороговому критерию, примененному к расстоянию $\rho(\mathbf{H}, \mathbf{w}_{\cdot k})$ с двумя небольшими отличиями от обычного RBF-нейрона. Первое — нейрон должен содержать значение своего порога — вещественное число, которое, вообще говоря, отлично от порогов других нейронов и должно настраиваться в процессе обучения. Оно характеризует радиус гиперсферы. Второе отличие состоит в том, что нейрон должен выдавать единичный сигнал в том случае, если вектор принадлежит внутренней области гиперсферы, т. е. когда значение на входе нейрона меньше (а не больше!) порога, что соответствует функции $s(x) = 1 - \vartheta(\lambda)$, где $x$ — вход, $\lambda$ — порог, а $\vartheta(x)$ — ступенчатая функция со скачком в нуле.

Неизбежность введения второго скрытого слоя вытекает как раз из того факта, что для описания класса требуется, возможно, более одного нейрона, также следует учесть порядок их следования и количество нейронов, отвечающее за каждый класс. Поэтому задачей этого, третьего по счету слоя

нейронов является обобщение результата, выдаваемого предыдущим слоем, и представление его в приемлемой форме.

В качестве четвертого, выходного слоя используется слой нейронов, соответствующий стандартной архитектуре Кохонена «победитель забирает все» [69], но победитель в данном случае определяется по минимуму расстояния, т. е. фактически по наибольшему соответствию одному из найденных классов. Полученная RBF-ИНС с метрикой $L_2$ и «соревновательным» (competive) алгоритмом выходного слоя ниже называется CRBFL2-ИНС. Значительную роль в успешной работе CRBFL2-сети сыграл предложенный для нее *алгоритм обучения*:

1. Первоначально количество радиальных нейронов третьего слоя устанавливается равным количеству ожидаемых классов. Вектор весовых коэффициентов нейрона приравнивается к произвольно выбранному элементу из класса, за который отвечает нейрон.

2. Выбирается случайный элемент $h_k$ из входного множества (т. е. выхода из первого скрытого слоя) и производится проверка условия: попадает ли этот элемент в одну из областей, описывающих данный класс. Если да, то соответствующий элемент помечается как больше не принимающий участия в процедуре выбора из данного пункта, и выбирается следующий случайный элемент.

3. Если нет, то радиус гиперсферы расширяется до возможности включить данный нейрон в область описания (внутрь гиперсферы).

4. Проверяется, попадают ли при этом элементы других классов в увеличившуюся область.

5. Если нет, то повторить процедуру с п. 2.

6. Если да, то увеличить третий слой на один нейрон. Весовой вектор выделенного нейрона приравнять к данному вектору входного множества, а порог — некоторой минимальной величине, гарантирующей попадание внутрь гиперсферы только одного данного входа.

7. Повторить с п. 2 до тех пор, пока множество, из которого разрешено выбирать вектора, не будет исчерпано.

Наиболее замечательной особенностью данного алгоритма является динамическое изменение числа нейронов в ходе обучения. Количество нейронов третьего слоя неизвестно заранее. Причина этого в том, что множество, которое небходимо описать, по своей форме, вообще говоря, отлично от гиперсферы и, следовательно, требует нескольких нейронов для описания.

Объектно-ориентированная программа на языке C++, реализующая предложенный алгоритм, тестировалась вначале на задаче распознавания искаженных символов алфавита. На рис. 16 приведены некоторые результаты распознавания как неискаженных, так и искаженных и зашумленных букв алфавита. После этого программа тестировалась на распознавании нескольких десятков человеческих лиц, причем после обучения на 3–4 снимках каждого

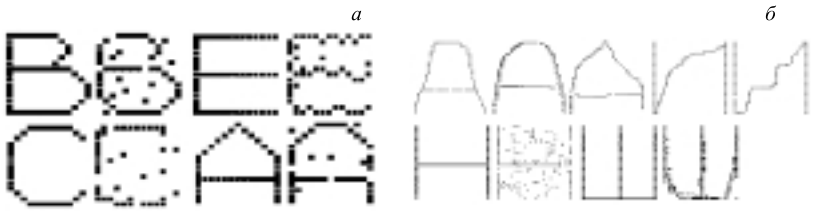



Рис. 16. Два примера букв: сначала идет обучающее изображение, за ним искаженное, но правильно распознанное программой: *а*) первый набор; *б*) второй набор, имитирующий ручное написание с зашумлением

из них в различных позах* показала 98 % надежность при распознавании их специально зашумленных и искаженных изображений. В качестве примера работы программы на рис. 17 приведены изображения лиц трех людей, различных по полу и возрасту, введенные в компьютер с видеокамеры как обучающее множество. Некоторые результаты распознавания этой группы лиц после внесения зашумлений или специальных искажений, существенно меняющих их облик, представлены на рис. 18. Все данные образы были правильно идентифицированы с породившими их лицами.

Рис. 17. Обучающее множество фронтальных изображений человеческих лиц

Однако при тестировании на обучающем множестве из 40 лиц, сфотографированных в десяти различных фронтальных позициях, полученных по Интернету с сайта Кембриджского университета [68], программа показала недопустимо высокий процент ошибочного «узнавания чужих» изображений. Основной причиной таких ошибок служила уже отмеченная выше в п. 2.1.2 ситуация, когда в пространстве входных признаков с размерностью 8000 (для изображений на растре $80 \times 100$) множество точек, описывающих признаки человеческого лица (даже после их усреднения в первом слое нейронов), занимает совершенно ничтожное подпространство много меньшей размерности. В этой связи были проведены предвари-

*Оцифровки изображений этих поз попросту усреднялись с помощью весов первого слоя нейронов.

тельные вычисления по статистической оценке ковариационной матрицы (45) для имевшихся 400 изображений лиц, после чего с помощью метода главных компонентов были вычислены коэффициенты ортогонального преобразования (46) в подпространство всего с 60 главными компонентами (на рис. 19 показан график спадания собственных значений ковариационной матрицы с ростом их номера, послуживший обоснованием для выбора размерности подпространства). Подстановка коэффициентов сокращающего МГК-преобразования в качестве весов первого слоя позволила без проблем реализовать его в рамках предложенной архитектуры CRBFL2-ИНС. Изображения, полученные с помощью МГК, представлены на рис. 20. Как видно, они оказались слишком зависимыми от вариаций изображений, от освещенности, фона и пр., что не улучшило характеристики распознавания. В этой связи пришлось прибегнуть к еще одному предварительному, «досетевому» двумерному вейвлет-преобразованию данных с помощью гауссовых вейвлетов второго порядка $g_2$ (известных как «мексиканская шляпа»). Детали вейвлет-преобразования описаны ниже в разд. 4, а здесь на рис. 20, *б* показаны результаты последовательного применения МГК к тем же изображениям, что на рис. 20, *а*, но предварительно преобразованных с помощью двумерных $g_2$-вейвлетов. Сравнение рис. 20, *а* и *б* показывает, как нормализовались вариации в освещенности и фоне.

Рис. 18. Зашумленные или специально искаженные лица из предыдущего набора, правильно распознанные CRBFL2-ИНС

Результаты распознавания на имеющемся тестовом наборе изображений с использованием двумерного вейвлет-преобразования с последующим применением МГК для сокращения числа входных признаков до 60 вполне удовлетворительны (эффективность на уровне 98,5 %) для практических приложений этого метода.

Наиболее очевидным достоинством подхода является, как показала практика, большая скорость обучения CRBFL2-ИНС. Также изначальная конечность процесса

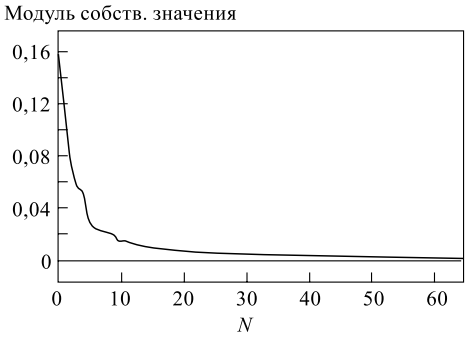



Рис. 19. Спадание величин собственных значений ковариационной матрицы в зависимости от их номера

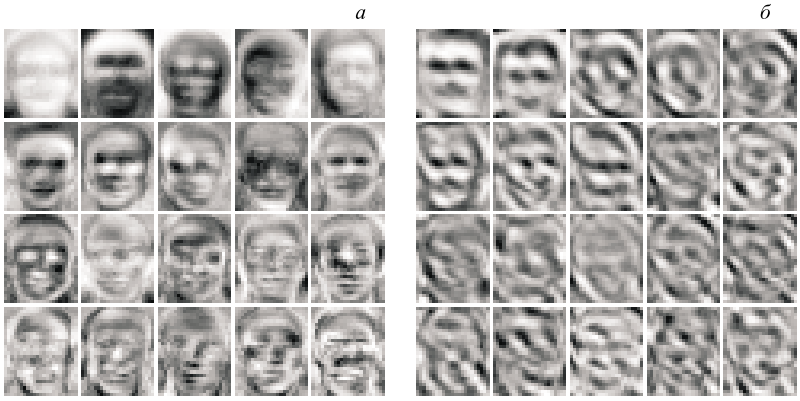



Рис. 20. *а*) Изображения некоторых лиц из кембриджской базы данных, полученные методом главных компонент. *б*) Результат применения МГК к лицам, предварительно подвергнутым двумерному вейвлет-преобразованию

обучения — несомненное преимущество перед методом обратного распространения ошибки, время обучения сети этим методом на многих задачах склонно затягиваться до практической бесконечности.

Возможность простой оценки качества классификации, проведенной CRBFL2-сетью в результате процесса обучения, — еще одно достоинство предлагаемого подхода. Критерием здесь служит количество радиальных нейронов после обучения сети. В том случае, когда оно равно количеству классов, это означает, что ИНС достигнута максимальная степень обобщения. Ситуация, когда количество нейронов в третьем слое совпадает с количеством элементов обучающего множества на его входе, означает, что обобщение не достигнуто из-за чрезвычайной разрозненности данных и сеть попросту «запомнила» все элементы.

При работе CRBFL2-ИНС в режиме распознавания может возникнуть ситуация, при которой входной вектор попадет в область, которую перекрывает несколько гиперсфер нейронов, принадлежащих разным классам. В таком случае нейронная сеть выдаст вероятностный ответ о принадлежности данного элемента к каждому из классов. Например, если вектор на входе сети вызвал единичный отклик у двух нейронов первого класса и одного нейрона второго класса, то, соответственно, вероятность его принадлежности к первому классу 2/3, а ко второму, соответственно, 1/3. Стоит также упомянуть, что если поданный на вход образ совершенно не похож ни на один из элементов обучающего множества, классифицированного сетью, то все нейроны выдадут нулевой сигнал, свидетельствующий о непринадлежности ни к какому из известных классов сети.

Проведение процедуры усечения, т. е. удаления связей, мало влияющих на конечный результат, также возможно для предложенной CRBFL2-ИНС. Поскольку каждый нейрон характеризуется радиусом гиперсферы, которую он описывает, то нейроны, у которых этот радиус мал, можно исключить из конечной сети, т. к. наиболее вероятно, что он описывает весьма немногочисленное количество элементов обучающего множества или попросту его случайный фрагмент.

Заметим также, что приведенные алгоритмы CRBFL2-ИНС вполне допускают возможность совершенствования предложенной двухуровневой схемы распознавания образов для реализации иных видов препроцессинга, либо в виде отдельной процедуры, либо в линейном случае в виде дополнительного слоя нейронов, работающего каким-либо специфическим образом.

Таким образом, сеть CRBFL2 благодаря своим замечательным способностям динамической настройки числа нейронов и их связей в процессе обучения и «умению» распознавать не только связные, но и несвязные подмножества в пространстве образов оказывается гораздо более похожей на реальные биологические нейросистемы в сравнении с обычными прямоточными ИНС, использующими метод ОРО для вычисления фиксированных весов нейронных связей.

Вспомнив о биологических нейросистемах, мы в заключение этого раздела отметим работу [70], посвященную исследованию динамики модели реалистичных нейронных сетей (RNN) в их реакции на внешние периодические несинусоидальные возбуждения разной формы. Численные эксперименты на базе уравнений динамики RNN позволили воспроизвести некоторые нейрофизиологические результаты [71] и обнаружить новые эффекты в поведении RNN.

**2.2. Применение полносвязных ИНС.** Как уже упоминалось, в общую схему ИНС, помимо широко применяемых персептронов, укладывается также сеть Хопфилда (ХНС) [72]. Это полносвязная сеть, являющаяся в простейшем случае [73] системой простых бинарных нейронов $s_i$, которые могут принимать одно из двух различных значений, например, $+1, -1$. Эволюция ХНС приводит ее в некоторое состояние стабильности, устойчивого равновесия. Рассматривая ИНС как динамическую систему бинарных нейронов, Хопфилд использовал билинейную функцию Ляпунова как функционал энергии сети

$$E(s) = -\frac{1}{2}\sum_{ij} s_i w_{ij} s_j \tag{48}$$

и показал, что для симметричной матрицы весов $w_{ij} = w_{ji}$ с нулевой диагональю $w_{ii} = 0$ и асинхронной динамикой сети ее энергия в результате эволюции убывает в локальные минимумы, соответствующие точкам стабильности сети.

В соответствии с вышесказанным для нахождения стационарного состояния сети требуется найти точку минимума энергетического функционала по значениям состояний нейронов. Применение с этой целью метода градиентного спуска к функционалу энергии (48) приводит к системе уравнений, определяющей динамику сети:

$$s_{ij} = \frac{1}{2}\left(1 + \operatorname{sign}\left(-\frac{\partial E}{\partial s_{ij}}\right)\right). \tag{49}$$

Однако процедура итерационного решения системы (49) для случая бинарных нейронов часто приводит в какой-нибудь локальный минимум функционала энергии. Кроме того, во многих практических приложениях ХНС бинарные нейроны оказываются нереалистичной идеализацией. В этой связи Хопфилд в 1984 г. [74] предложил обобщение своей модели ИНС на случай нейронов с непрерывным множеством состояний.

Стандартным путем перехода к нейронам с непрерывными состояниями является введение статистического шума в систему с последующим применением теории среднего поля (см., например, [40]), что приводит к усреднению значений состояний нейронов и замене ступенчатой функции (49) на функцию сигмоидального вида, аналогичную (38) при $\lambda = 1/T$:

$$v_{ij} = \frac{1}{2}\left(1 + \tanh\left(-\frac{\partial E}{\partial v_{ij}}\frac{1}{T}\right)\right). \tag{50}$$

Здесь температура $T$ соответствует уровню статистического шума. В соответствии с теорией среднего поля состояния нейронов $v_{ij} = \langle s_{ij}\rangle_T$ усредняются по температуре, и значения дискретных нейронов становятся уже непрерывными в области [0, 1]. С помощью системы уравнений (50) состояния нейронов сети итеративно обновляются до достижения ею стабильной точки. Однако то, что нейроны теперь не являются бинарными, позволяет оперативно следить за теми из них, которые соединяют точки треков и под стимулирующим воздействием весов постепенно увеличивают уровень своей активности в процессе эволюции сети. В качестве порога уровня активности $v_{ij}$ обычно выбирается $v_{\min} = 0,1$.

Работы Хопфилда породили чрезвычайный интерес к таким сетям, в частности, потому, что устанавливалась явная связь процессов их эволюции с обширным кругом проблем оптимизации, формализуемых обычно как задача поиска экстремума функционала при наличии ограничений на его параметры.

К таким задачам относится и задача поиска треков по данным координатных измерений в современных электронных детекторах типа пропорциональных, дрейфовых или стриповых камер. Если определять связи между нейронами сети, исходя из измеренных значений координат, то это приводит

к чрезмерно большому числу нейронов и еще большему числу межнейронных связей, подавляющее большинство которых не относилось к трекам. В итоге любая программная реализация сети оказывалась слишком громоздкой и работала настолько медленно, что становилась бесполезной в практических приложениях. Для борьбы с этим «проклятием размерности» весовые функции вводились так, чтобы поощрять связи нейронов, принадлежащих одному и тому же треку, а в функционал энергии вводились «ограничивающие» члены, запрещающие как межтрековые связи, так и чрезмерный рост числа самих треков.

Первую удачную попытку использовать нейронные сети для распознавания треков сделали Петерсон и Денби [75–77]. Их подход (так называемый *метод сегментов*) был подробно описан в обзоре [40], так что мы здесь только укажем, что на множестве экспериментальных точек на плоскости вводились нейроны $v_{ij}$, определяющие, соединяются точки $i$ и $j$ или нет, т. е. принадлежит данный направленный сегмент $v_{ij}$ треку или нет. В дальнейших вычислениях в соответствии со схемой (50) состояния нейронов перестают быть целочисленными. Их значения и определяют уровень активности нейрона, т. е. в случае $v_{ij} > v_{\min}$ нейрон считается активным. Энергетический функционал в работах [75–77] был определен как состоящий из двух частей:

$$E = E_{\text{cost}} + E_{\text{constraint}}. \tag{51}$$

Предполагалось, что распознавание велось для гладких прямых треков без разветвлений. Поэтому первый стоимостный член выбирался так, чтобы он поощрял *короткие смежные сегменты с малым углом* между ними, а второй член (штрафной) устанавливал *запрет разветвлений* (т. е. запрет ситуаций, когда к одной точке присоединено больше одного сегмента-нейрона) и *баланс между числом активных нейронов и числом экспериментальных точек*.

К сожалению, эта схема Денби–Петерсона, принципиально запрещающая бифуркацию трека, была неприменима в случаях распадов нейтральных и рождения заряженных частиц в объеме детектора. Поэтому в ходе работы [79], в которой исследовались данные распада нейтральных каонов и гиперонов, пришлось модифицировать штрафной член энергетической функции так, чтобы он позволял осуществлять распознавание разветвлений. Для соответствующего смягчения этой штрафной части она использовалась только тогда, когда количество активных нейронов, имеющих общую точку с данным нейроном, превышало четыре, а в остальных случаях бралась только ее половина. Пример работы сети дан на рис. 21, где реальные треки показаны сплошной линией, а ложные связи отмечены пунктиром. В тяжелых фоновых условиях и при экспериментальной неэффективности камер действующая программа обработки эксперимента EXCHARM не распознала 5,2 % событий, а программа, основанная на ИНС, 3,2 %. При этом из нераспознанных событий для разных программ совпали только 0,8 % от общего числа событий,

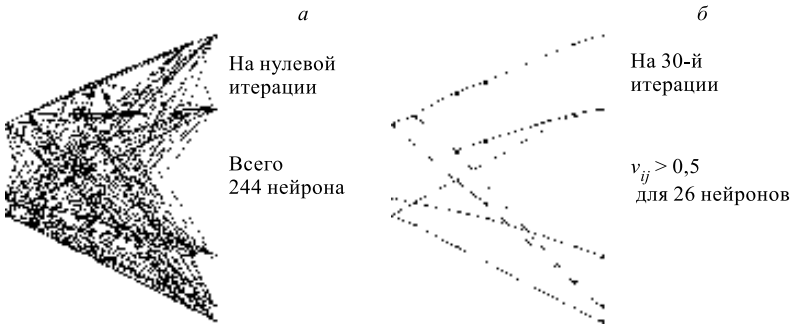



Рис. 21. Исходное (*а*) и конечное (*б*) состояния сети

т. е. эти два множества плохо распознанных событий имеют малую область пересечения. Таким образом, возникла идея эффективного комбинированного алгоритма восстановления треков при незначительном увеличении временных затрат на обработку. Идея заключается в следующем: после распознавания $x$- и $y$-проекций треков более быстрым методом опорных дорожек в программе обработки производится пространственная «сшивка» этих проекций. Если часть надежно найденных проекций не имеет сшивок, то включается нейросетевой алгоритм, который в большинстве случаев успешно решает задачу распознавания треков в таких событиях.

Комбинированный алгоритм с использованием ИНС был опробован и на реальных данных эксперимента EXCHARM и позволил достичь 99 % эффективности распознавания событий.

В обзорах [40, 80] также было дано подробное описание метода *роторных ИНС с префильтрацией данных с помощью клеточных автоматов* и его применение для распознавания треков в магнитном поле. Этот метод был с успехом применен для распознавания треков по данным, полученным на спектрометре ARES [78]. Напомним, что ротором-нейроном в этой работе назван вектор $\mathbf{v}_i$, исходящий из точки, измеренной на треке, представляющем дугу окружности. Направление ротора должно в идеале идти по касательной к треку, а модуль должен характеризовать интенсивность воздействия этого нейрона на остальные. Как показано в [78], энергетический функционал такой сети выражается общей формулой (48) с весами

$$w_{ij} = \begin{pmatrix} \cos 2\varphi_{ij} & \sin 2\varphi_{ij} \\ \sin 2\varphi_{ij} & -\cos 2\varphi_{ij} \end{pmatrix}, \tag{52}$$

где $\varphi_{ij}$ — угол между хордой $r_{ij}$, соединяющей начальные точки $i$-го и $j$-го роторов, и осью 0X.

Мы укажем здесь на два направления, по которым пошло далее развитие этого подхода:

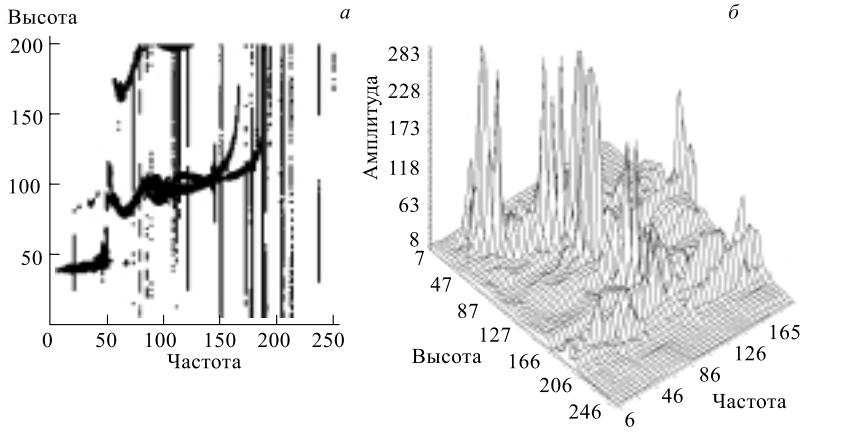



Рис. 22. Двумерное (*а*) и трехмерное изображения ионограммы с учетом амплитуды сигнала (*б*)

- модификация роторных ИНС для обработки данных ионограмм;
- применение клеточных автоматов (КА).

*2.2.1. Анализ ионограмм.* Данные, полученные в результате вертикального зондирования ионосферы, представляют собой последовательность отсчетов частоты сигнала, его амплитуды и времени прихода после отражения от ионосферного слоя [81, 84]. Типичная ионограмма вертикального зондирования показана на рис. 22, *а*, где по оси абсцисс дана измеренная частота сигнала, по оси ординат — высота слоя. Расплывчатые кривые, сформированные отсчетами, так же, как в ФВЭ, называются треками. Ионосферные треки могут пересекаться и искажаться вертикальными помехами. Если учесть еще и амплитуды приходящих отсчетов, то получим трехмерную картинку, изображенную на рис. 22, *б*, напоминающую горный пейзаж. Задача состоит в том, чтобы с минимальными временными затратами без использования априорной информации распознать ионосферные треки и аппроксимировать их по точкам наибольшей амплитуды, т. е., по альпинистской терминологии, зафиксировать траверсы всех «хребтов». Эта задача была успешно решена в [81] методом модифицированных роторных ХНС с предварительной обработкой данных с помощью клеточных автоматов. Основной целью этой предобработки было устранение вертикальных помех и заполнение пропущенных отсчетов. После этого для каждого частотного кластера находилась наибольшая амплитуда. После предобработки, сокращавшей объем данных в 5–10 раз, данные поступали в роторную ИНС, которую пришлось модифицировать, чтобы справиться с переменной кривизной треков ионов. Модификация состояла в отказе от глобальной аппроксимации трека окружностью и введении скользя-

щего окна просмотра, размер которого выбирался так, чтобы в его пределах локальная аппроксимация окружностью выполнялась с заданной точностью. Идея скользящего окна упростила и ускорила также процедуру определения начальной конфигурации сети путем углового гистограммирования (см. [78]), ограничив ее пределами текущего окна. Таким образом, инициализация по сжатым данным выполнялась следующим образом: каждая точка помещалась в середину левой стороны окна, разбитого на угловые сектора в $1^\circ$. Подсчитывалось число точек в пределах окна, попавших в каждый сектор. Центр тяжести трех смежных ячеек с максимумом этой гистограммы в середине и брался в качестве направления ротора, а за модуль принималось среднее от числа точек в этих трех ячейках. Динамика среднего поля для эволюции ротора определялась известной формулой [40, 80]:

$$v_i = \tanh\left(\sum_{j=1} w_{ij} v_j / T\right) \tag{53}$$

с $T = 1,5$. Процедура сходилась в среднем за 3–5 итераций, что подтвердило ее эффективность, а также оправдало значительные временные затраты на инициализацию роторов, занимавшую иногда до 90 % времени обработки. В дальнейшем этот метод и реализующая его программа активно использовались в реальных экспериментальных исследованиях ионосферы в России и США [82–84].

*2.2.2. Применение клеточных автоматов.* Трактуя нейронные сети как дискретные динамические системы, можно и клеточные автоматы рассматривать как вариант такой простой нейронной сети, в которой нейроны-клетки принимают лишь конечное дискретное множество состояний (часто всего два), их связи локализованы ближайшими соседями, правила эволюции также жестко определены в зависимости от них, и, наконец, сама эволюция происходит в тактовые моменты времени, синхронные для всех клеток автомата. Тем не менее эта кажущаяся простота структуры и динамики позволяет организовать самые удивительные и полезные применения КА. В соответствии с темой данного обзора мы ограничимся только применениями КА в ФВЭ, отсылая читателя к имеющейся литературе по поводу иных приложений [101, 102].

Правила построения клеточного автомата для отсева шумовых отсчетов, заполнения пропусков и распознавания связных групп точек в процессе эволюции КА были предложены в [99] как обобщение правил известной игры «жизнь» [103]. В дальнейшем эти правила применялись с необходимыми модификациями для предварительного грубого распознавания сегментов треков с последующим более точным распознаванием треков или вершин событий путем применения ИНС.

Эти модификации были необходимы для введения критерия близости клеток по направлению вдоль трека. Подобно тому, как это делалось в сегментной модели ИНС Денби–Петерсона [75–77], клетка определялась как сегмент, соединяющий два отсчета на соседних координатных плоскостях (с учетом неэффективности камер допускались соединения и со следующей камерой). Соседство устанавливалось по близости сегментов по направлению, т. е. запрещались большие углы (хотя возможность многократного рассеяния могла допускаться). При инициализации значения клеток с допустимыми наклонами сегментов устанавливались в единицу, затем в процессе эволюции значение клетки увеличивалось на единицу, если у соседа на предыдущем слое было то же значение. Эволюция прекращалась, когда не оставалось клеток с одинаковыми значениями. После этого треки собирались, начиная с клетки с наибольшим значением, путем присоединения соседей с предыдущими значениями при движении назад вдоль дерева. Если появлялись лишние ветви, оставлялись более длинные и гладкие. На заключительной стадии проверялось качество треков для отсева ложных из них по критерию $\chi^2$ и большему числу сегментов [107].

В работах [104, 105] в экспериментах DISTO и STREAMER такой КА применялся вначале для распознавания прямолинейных треков, после чего для поиска вторичной вершины и идентификации обнаруженных частиц использовались ИНС типа МСП. В работе [107] КА использовались в эксперименте NEMO-2 для поиска сегментов треков, которые затем подгонялись методом эластичных ИНС (см. разд. 3). В недавней публикации [108] приведены данные об успешном применении программы CATS, реализующей КА для обработки треков в вершинном детекторе установки HERA-B.

В работах [86, 109] удалось сконструировать клеточный автомат совершенно другой структуры для генерации случайных многомерных векторов, что необходимо при решении многих актуальных задач физического моделирования методом Монте-Карло (см., например, [112]). В [86] рассмотрен двумерный бинарный клеточный автомат (ДБКА) на растре из $N \times 31$ клеток, свернутом в тор вдоль стороны с $N$ клетками. Сравнивались два правила эволюции ДБКА: ячейка $a_{ij}$ принимает значение, равное:

1) сумме по $\mathrm{mod}\,2$ значений восьми соседей, окружающих эту ячейку, т. е.

$$a'_{ij} = \left( \sum_{i'=i-1}^{i+1} \sum_{j'=j-1}^{j+1} a_{i'j'} \right) \mathrm{mod}\ 2; \tag{54}$$

2) той же сумме, но с добавкой $a_{ij} * a_{i+1,j+1} \mathrm{mod}\ 2$, т. е.

$$a'_{ij} = \left( \sum_{i'=i-1}^{i+1} \sum_{j'=j-1}^{j+1} a_{i'j'} \ + \ a_{ij} * a_{i+1,j+1} \right) \mathrm{mod}\ 2. \tag{55}$$

При правильно выбранном алгоритме эволюции ДБКА после ее завершения каждая из $N$ строк обновленной матрицы после ее деления на $2^{31}-1$ может быть использована как случайное число, равномерно распределенное в интервале (0, 1), т. е. фактически получается случайный вектор с $N$ компонентами. Статистическое поведение полученных по этим двум правилам двух генераторов случайных векторов (ГСВ) сравнивалось по таким тестам, как длина периода, наличие корреляции между случайными числами и их распределение в пространствах размерности до 20. ГСВ проверялись также и для разного числа компонентов. Качество многомерных распределений проверялось по оригинальному методу вложенных гистограмм [109]. В результате проведенных тестов выяснилось, что полученные с помощью обоих ГСВ случайные последовательности некоррелированы и равномерно распределены в многомерных единичных кубах. Тест на длину периода показал преимущество второго из генераторов (ГСВ2), имеющего период, заведомо превышающий $10^8$ при $N = 2, 4, 5, \ldots, 10, 25, 50, 100$. Дополнительно ГСВ2 показал положительный результат при проверке по тестам, предложенным в известной книге Д. Кнута [110]. Наиболее существенной оказалась проверка ГСВ2 в задаче вычисления теплоемкости в точке кроссовера для $SU_3$-калибровочной теории на решетке, показавшая, в частности, отсутствие корреляции на расстоянии порядка 5000 [111].

*2.2.3. Управляемые нейронные сети.* Решение широкого класса комбинаторных задач, объем вычислений которых растет как экспонента с размерностью задачи, можно выполнять в разумное время с помощью особого типа нейронных сетей, названных в работе [87] *управляемыми ИНС*. Одной из классических проблем комбинаторной оптимизации, относящейся к классу $NP$-проблем, т. е. задач, не решаемых за число шагов, растущее как полином от $n$, является, например, задача о расстановке $N$ ферзей на шахматной доске размером $N \times N$ так, чтобы они не били друг друга. Программы, основанные на обычных комбинаторных методах, уже при $N \geq 97$ не могут решить эту задачу за сколько-нибудь реальное время, в то время как управляемая ИНС, построенная на базе сети Хопфилда, позволила решить эту задачу для $N = 1024$ за 10 минут на компьютере 90-х годов.

Идея управляемой ИНС (УИНС) заключается во введении в сеть дополнительных нейронов, называемых «контролерами», предназначенных для управления эволюцией определенных групп нейронов с целью предотвратить «застревания» энергетической функции сети в локальном минимуме. Выбор групп управляемых нейронов зависит от специфики задачи, в частности, в задаче расстановки ферзей в качестве группы выбирается одна из сторон шахматной доски.

В ФВЭ задача сшивки треков, распознанных на разных проекциях (track-match) является именно такой комбинаторной $NP$-проблемой. В терминологии теории графов задача сшивки треков сводится к построению матрицы

*Таблица* 2. **Время решения задачи сшивки треков в зависимости от их числа**

| $N$ | 30 | 15 | 13 |
|---|---|---|---|
| УИНС | 105,32 с | 0,26 с | 0,13 с |
| САП | 7,5 ч | 23,78 с | 0,135 с |

инцидентности с выбором ненулевого элемента на каждой ее строке так, чтобы все эти элементы лежали в разных столбцах. На шахматном языке эта задача означает, что мы должны на шахматной доске размером $N \times N$ поставить более чем $N$ ладей, не оставив ни одной пустой строки, а потом снять «лишние» ладьи так, чтобы на каждой строке оставалось ровно по одной ладье и ни одна не попадала под удар другой. Описание деталей конфигурации соответствующей УИНС и определения весовой матрицы, обеспечивающей функционирование нейронов-контролеров, чересчур громоздки для изложения в рамках данной статьи. Их можно найти в [87]. Мы приведем здесь результат сравнительного прогона программы, реализующей УИНС, и программы, основанной на быстром стандарном алгоритме перебора (САП) (см. табл. 2). Вычисления велись на VAX-8350.

Как видно, начиная с $N \geq 30$, УИНС продолжает показывать разумное время, тогда как применение стандартного алгоритма становится нецелесообразным.

Алгоритм УИНС может быть адаптирован для решения и других задач, например, составления расписаний больших размерностей.

## 3. МЕТОД ЭЛАСТИЧНЫХ НЕЙРОННЫХ СЕТЕЙ ИЛИ ГИБКИХ ШАБЛОНОВ

Как отмечалось во введении, экспериментальную ФВЭ характеризует активный рост энергий пучков частиц и загрузки детекторов, ведущий к возрастанию множественности взаимодействий и, соответственно, новых требований к методам распознавания событий. Это особенно верно для экспериментов с ультрарелятивистскими взаимодействиями тяжелых ионов, таких, например, как на ускорителе RHIC в BNL, где множественность взаимодействий уже достигает значений в тысячи треков, а на новом коллайдере LHC в CERN ожидаются множественности в десятки тысяч треков. Более того, требования к точности определения вершины взаимодействия и обнаружения короткоживущих частиц вынуждают физиков размещать детекторы в нескольких сантиметрах от центральной области взаимодействий, где высочайшая загрузка и плотность отсчетов даже при предельно точных высокогранулярных

детекторах делают невозможным применение традиционных методов прослеживания треков или наведения по дорожке. Еще в 1991 г. Джиласси и Харландер (ДХ), озабоченные наступлением подобных проблем, провели известное сравнительное исследование трех методов трекинга: классического — наведения в дорожке (RF — road finder), нового тогда еще метода нейронных сетей (NN — neural networks) и предложенного ими эластичного трекинга (ET) [92]. Критерием для сравнения по плотности загрузки они предложили *трековую плотность* $\rho_{\mathrm{track}}$, определенную как среднее отношение расстояний между точками, измеренными вдоль трека, и точками, принадлежащими разным трекам или шумовым точкам. В их работе [92] приведены убедительные данные, свидетельствующие о том, что если для $\rho_{\mathrm{track}} \leq 1$ RF- и NN-методы трекинга дают вполне удовлетворительные результаты, то уже при $1 \leq \rho_{\mathrm{track}} \leq 4$ RF-метод перестает надежно работать и только NN-метод выдерживает конкуренцию с ET-трекингом. При $\rho_{\mathrm{track}} \geq 5$ и нейросетевые методы быстро перестают быть эффективными, так что только метод ET может справиться с задачей трекинга в условиях столь высоких загрузок и зашумленности вплоть до $\rho_{\mathrm{track}} \sim 10$, что соответствует загрузкам, ожидающимся на LHC.

Прежде чем объяснить, что же такое метод эластичного трекинга, следует вспомнить, что термин «эластичные нейронные сети» был ранее введен Дурбином и Вилшоу (ДВ) [93] при решении известной «задачи коммивояжера», который должен объехать $n$ городов кратчайшим путем, не заезжая ни в один из них дважды. Дурбин и Вилшоу начали с того, что в центр плоскости с размещенными на ней городами поместили маленький эластичный замкнутый контур (окружность) с $n$ нейронами на нем и заставили затем этот контур растягиваться путем приложения к каждому из нейронов двух типов сил: одна двигает его к ближайшему из городов, другая отталкивает его от соседей на контуре.

Эти силы заставляют контур постепенно растягиваться в ходе итеративных перевычислений позиций нейронов, пока вся эта эластичная сеть не установится, образуя оптимальный обход.

Подход ДВ был удачно обобщен авторами [94, 106, 107] при обработке данных с дрейфовых камер установки NEMO с их право-лево-неопределенностью и наличием многократного рассеяния, ведущего к изломам треков. Как указывалось в п. 1.3.3, при проводке треков в дрейфовых камерах следует помнить, что треки должны касаться маленьких окружностей, образованных радиусами дрейфа. Считая обе возможные точки касания вычисленными по координатам центральных проволок и известным радиусам дрейфа, авторы работы [94] взяли эти точки в качестве городов в схеме ДВ, а для устранения право-лево-неопределенности добавили еще *третью силу*: притяжение между этими точками, выдавливающее геометрическую область внутри контура так, чтобы стянуть его в реальный трек (см. рис. 23, взятый из [94]).

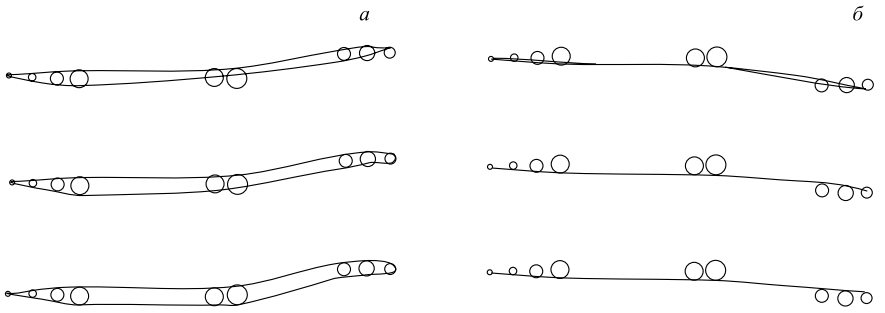



Рис. 23. Пример эволюции эластичной нейросети: начало (*а*) и окончание (*б*) эволюции

Предлагая метод эластичного трекинга, Джиласси и Харландер исходили из другой идеи, не связанной с ИНС (хотя в [92] они показали, что можно трактовать ЕТ-метод как один из вариантов ИНС Хопфилда). Джиласси и Харландер исходили из идеи гибкого шаблона, т. е. уравнения трека, зависящего от вариаций параметров таким образом, чтобы, изгибаясь при их изменении, кривая, описываемая этим уравнением (ее еще называют деформируемым шаблоном, эластичной рукой и т. д.), прошла как можно ближе к точкам, измеренным на треке. Этот подход можно с физической точки зрения описать как взаимодействие 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

$$E = -\int dr' dr \rho_T(r) V(r - r') \rho(r') \;\longrightarrow \min, \tag{56}$$

где $V$ — потенциал, зависящий от расстояния измеренных точек до шаблона. Джиласси и Харландер выбрали потенциал Лоренца

$$V(x,t) = \frac{w^2(t)}{x^2 + w^2(t)} \tag{57}$$

с шириной, зависящей от времени:

$$w(t) = b + (a - b)\,\mathrm{e}^{-t/T}, \tag{58}$$

где $T$ — временная константа; $a$ — максимальное расстояние, на котором точки еще приписываются к данному шаблону; $b \approx \sigma_{\mathrm{res}}$ — точность про-

странственного разрешения детектора. Очевидно, что $b \ll a$. Учитывая дискретность измерений, получаем вместо интеграла в (56) сумму

$$E(\pi, t) = -\frac{1}{N} \sum_{i}^{N} \frac{w^2(t)}{(\mathbf{x}_i - \mathbf{r}(\pi, \mathbf{x}))^2 + w^2(t)} \ . \tag{59}$$

Здесь $N$ — число точек на треке; $\mathbf{x}_i$ и $\mathbf{r}$ — $i$-я точка, измеренная в пространстве, и ее расстояние до шаблона; $\pi$ — набор параметров трека. Если рассматривать детектор в однородном магнитном поле, то уравнением трека будет винтовая кривая (геликоида) со следующими параметрами в точке $\mathbf{x}_i$: кривизна $\kappa$, угол погружения $\lambda$ и фаза $\Phi$, т. е. $\pi = \{\kappa, \lambda, \Phi, \mathbf{x}_0\}$.

Функционал в (59) зависит от точек только одного трека, хотя в принципе можно осуществлять одновременную подгонку сразу всех треков. Тем не менее ДХ этого не рекомендуют. Наличие зависимости потенциала от времени позволяет применять технику имитированного отжига (см. выше в разд. 1.1): вначале, когда параметры известны только в грубом приближении, берется потенциал настолько широким, чтобы наверняка захватывать все измеренные точки, с тем чтобы потом по мере уточнения параметров сужать потенциал.

ЕТ-метод в такой постановке был успешно применен в [95] для распознавания треков по данным времяпроекционной камеры установки STAR в условиях, предельных по загрузке: до 5 тыс. треков на событие ($\rho_{\mathrm{track}} \geq 7$). Минимизация энергетического функционала (59) по параметрам $\pi = \{\kappa, \lambda, \Phi, \mathbf{x}_0\}$ велась методом наискорейшего спуска:

$$\Delta\pi_i(t) = -\eta \bigtriangledown_\pi E(\pi, t) + \alpha \Delta\pi_i(t-1) \tag{60}$$

с регулятором шагов по параметрам $\eta$ и инерционным членом для подавления флуктуаций при подходе к минимуму.

Весьма ответственная часть ЕТ-метода — это выбор начального приближения. В [95] для этого применен вариант локального преобразования Хафа (см. п. 1.3.3). Хотя выбор трех пространственных точек позволяет определить параметры геликоиды, получаемой при этом точности в измерении кривизны $\kappa$ оказалось недостаточно для определения надежно продолжаемого шаблона в данных тяжелых условиях по трековой плотности. Исследование, проведенное для выяснения, сколько именно точек в начальном шаблоне необходимо для обеспечения такой продолжаемости, показало, что требуется нарастить начальный триплет еще на три точки в смежных координатных плоскостях.

Наиболее времяемкой частью программы C++, реализующей ЕТ-метод для применения в реальных условиях, оказалась проблема быстрых манипуляций с сырыми данными без замедляющих процедур предварительной обработки. Для этого были развиты специальные программные средства типа Pixel

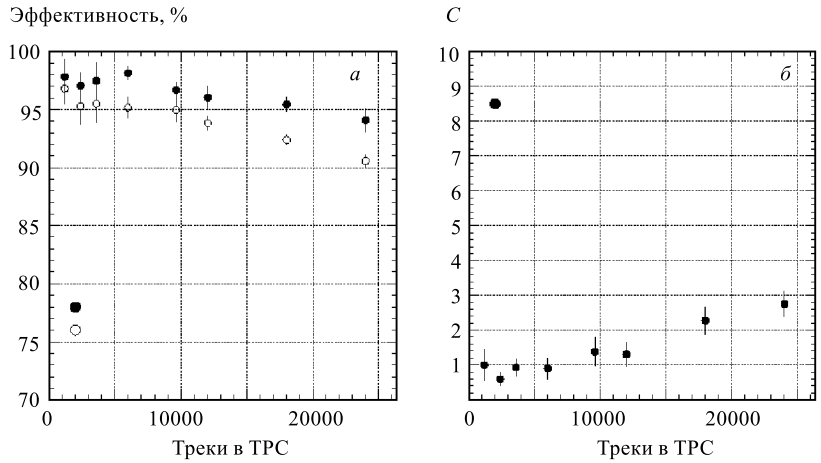



Рис. 24. Эффективность ($а$) и засорение $C$ ложными треками ($б$) в зависимости от множественности событий для ET-метода в применении к данным TPC STAR

Container [95]. В итоге удалось добиться требуемой эффективности распознавания треков (см. рис. 24) при заданной точности определения их параметров, хотя общие временные затраты на распознавание и определение параметров оказались довольно большими ($\sim$ 10 мин/событие с 2 тыс. треков на HP K200/RISC PA-8000 при дальнейшем квадратичном росте в зависимости от множественности).

Еще один подобный подход, названный методом гибких шаблонов (или деформируемых образцов — deformable templates) был предложен физиками Лундского университета Олссоном и Петерсоном (ОП). В отличие от функционала (59) в ДХ-подходе, они предложили глобализовать задачу, используя энергетическую функцию, учитывающую влияние всех $M$ треков события:

$$E(S_{ia},\pi)=\sum_{i}^{N}\sum_{a}^{M}S_{ia}D_{ia}(\pi,\mathbf{x})+\lambda\sum_{i}^{N}\left(\sum_{a}^{M}S_{ia}\right)^{2}, \qquad (61)$$

где $\pi$ — по-прежнему параметры геликоиды; $S_{ia}$ — бинарный нейрон, определяющий принадлежность $i$-й точки к $a$-му треку, т. е. в случае принадлежности $S_{ia}=1$, а в противном случае $S_{ia}=0$. $D_{ia}(\pi,\mathbf{x})$ — квадрат расстояния от точки до трека. Минимизация (61) ведется при условии, что каждая точка может принадлежать только одному треку или не принадлежать ни одному. В последнем случае минимизируемый функционал штрафуется на величину $\lambda$, 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 ведутся в соответствии с обычной схемой ХНС: сеть термализуется и применяется теория среднего поля. Для последовательности уменьшающихся температур метод наискорейшего спуска дает пошаговый итерационный алгоритм поиска минимума [96]:

$$\Delta\pi_a^{(k)} = -\eta_a^{(k)} \sum_i V_{ia} \frac{\partial D_{ia}}{\partial \pi_a^{(k)}} \ , \tag{62}$$

где для $\beta = 1/T$ потенциал $V_{ia}$, называемый фактором Поттса, определяется как

$$V_{ia} = \frac{\mathrm{e}^{-\beta\, D_{ia}}}{\mathrm{e}^{-\beta\, \lambda} + \sum_{b=1}^{M} \mathrm{e}^{-\beta\, D_{ib}}}. \tag{63}$$

В качестве инициализирующей процедуры ОП рекомендуют локальный вариант преобразования Хафа.

Метод деформируемых образцов был успешно применен в работах [26, 97] для обработки данных, параметризуемых уравнением окружности: колец черенковского излучения и треков в однородном магнитном поле.

В работе [97] ОП-методом осуществлялся поиск черенковских колец с одновременной оценкой их параметров по данным RICH типа CERES (см. пример на рис. 7). Как и в [16], исследование велось на модели данных, преобразованных в «хиты», т. е. после предварительной кластеризации и нахождения центров кластеров, но поиск велся глобальный и без всяких априорных сведений о центрах и радиусах колец. Эта информация была получена последовательной версией преобразования Хафа, начиная с перебора допустимых триплетов (35) и последующего двукратного гистограммирования (см. разд. 1.3.2). В [97] можно найти полезные рекомендации по выбору размеров биннинга гистограмм и порогов для оценки параметров колец, полученные на основе исследования о влиянии на эффективность распознавания таких факторов, как радиальный разброс хитов, общая зашумленность события и среднее число фотонов на кольцо. Об эффективности полученного алгоритма реализации ОП-метода свидетельствует то, что он с минимальной модификацией был также успешно применен для распознавания треков в магнитном поле. Примеры работы алгоритма в этих двух режимах представлены на рис. 25. Следует, однако, сделать важное замечание, касающееся издержек глобальности ОП-метода: 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.

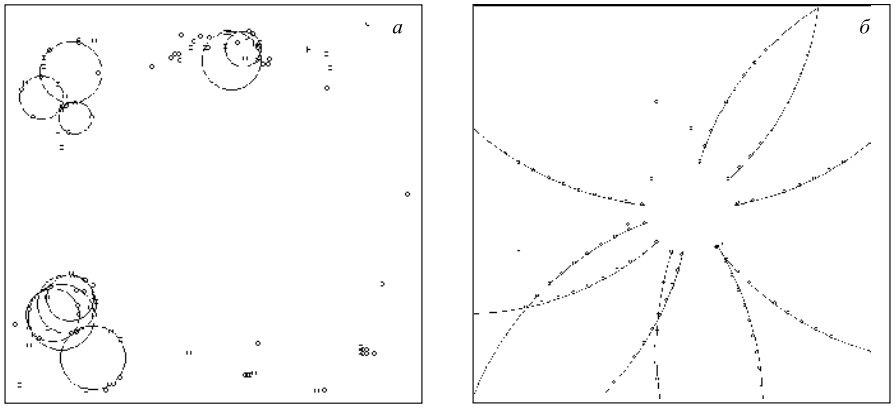



Рис. 25. Примеры результатов работы метода деформируемых образцов [97]: распознавание черенковских колец (*а*) и треков в магнитном поле (*б*)

Вторая работа [26] касалась применения ОП-метода в задаче распознавания и определения параметров треков в системе дрейфовых трубок, описанных выше в разд. 1.3.3, где результаты измерений в проекции $X0Z$ являются тройками чисел $\{x_i, z_i; r_i\}, i = \overline{1,\, N}$, где $(x_i, z_i)$ — координаты центра сработавших трубок; $r_i$ — радиус дрейфа. Считая трек окружностью с параметрами $(a, b; R)$, получаем для расстояния от центра $i$-й трубки до трека следующее выражение:

$$D_i(a, b; R) = R - \sqrt{(x_i - a)^2 + (z_i - b)^2},$$

которое, отражая суть проблемы право-лево-неопределенности, может быть как положительным, так и отрицательным. Обозначим

$$\begin{aligned} d_i^- &= (D_i(a, b; R) - r_i)^2, \text{ если } D_i(a, b; R) > 0, \\ d_i^+ &= (D_i(a, b; R) + r_i)^2 \text{ в остальных случаях.} \end{aligned} \tag{64}$$

Из-за этой двойственности пришлось вместо бинарных нейронов $S_{ia}$, используемых для определения принадлежности точки треку, ввести двумерный вектор $\mathbf{s}_i = (s_i^+, s_i^-)$ с допустимыми значениями $(1, 0), (0, 1), (0, 0)$. Функционал зависит теперь от пяти параметров:

$$E(a, b, R, s_i^-, s_i^+) = \sum_{i=1}^{N} \left\{ d_i^- s_i^- + d_i^+ s_i^+ + \lambda((s_i^- + s_i^+) - 1)^2 \right\}, \tag{65}$$

где $\lambda$ — ошибка измерения радиуса дрейфа, и минимизируется при ограничении $s_i^+ + s_i^- \le 1$. Факторов Поттса стало два:

$$s_i^- = \frac{1}{1 + \mathrm{e}^{(d_i^- - \lambda)/T} + \mathrm{e}^{(d_i^- - d_i^+)/T}}, \qquad s_i^+ = \frac{1}{1 + \mathrm{e}^{(d_i^+ - \lambda)/T} + \mathrm{e}^{(d_i^+ - d_i^-)/T}}.$$

При трехступенчатой последовательности убывающих температур, используемых для реализации схемы имитационного отжига, процедура минимизации оказалась достаточно изощренной и потребовала последующей корректировки параметров (см. детали в [26]). Однако, как отмечалось выше в разд. 1.3.3, временные затраты на метод деформируемых образцов удалось компенсировать, благодаря организации гибридной структуры обработки, в которой этот метод используется только в редких случаях.

В заключение этого раздела обратим внимание читателей на поразительную общность не только двух методов эластичных шаблонов или образцов, которые, несмотря на разницу в выводе формул для функционалов энергий, в итоге отличаются только выбором потенциалов, но на их общность с робастным подходом, описанным в разд. 1.1. *Формула (1б) для робастного функционала правдоподобия, по сути, описывает все варианты эластичных потенциалов*. Предложенный в разд. 1.2 способ определения оптимальной весовой функции методом максимального правдоподобия [6] может быть обобщен для применения при иных распределениях фоновых процессов, предоставляя строго обоснованный статистический путь вывода того или иного эластичного потенциала. Это, в частности, подтверждает и совпадение идеи о возможности обработки «сырых» данных, предложенное в работе [113] Джиласси и Харландером, а затем в работе [8], где дается более общий подход с использованием двумерной весовой функции (31). Отметим также работу [114], где робастный подход с гауссовым потенциалом применен для подгонки эллипсов по неполным и сильно зашумленным данным. Минимизацию сильно нелинейного функционала вида (16) удалось эффективно осуществить с помощью программы FUMIVI [115].

## 4. СВОЙСТВА ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЙ И ИХ ПРИМЕНЕНИЕ ДЛЯ АНАЛИЗА ДАННЫХ В ФВЭ

Фурье-анализ, ставший одним из основных методов машинного анализа экспериментальных сигналов после введения быстрого алгоритма дискретного преобразования Фурье, около 10 лет тому назад начал быстро уступать свои позиции вейвлет-анализу. Термин *вейвлет* («wavelet» (англ.) — маленькая волна) не сразу утвердился в русскоязычной научной литературе, и еще три года назад ряд известных исследователей предпочитали ему русские термины типа «всплеск» или «волнолет» [116], которые, однако, не прижились. Эффективность вейвлет-анализа в сравнении с преобразованием Фурье объясняется большей информативностью первого, который предоставляет исследователю дополнительную степень свободы для анализа в виде возможности видеть разложение сигналов по солитоноподобным базисным функциям при различных масштабах.

Одномерное вейвлет-преобразование сигнала $f(x)$ имеет следующий вид [117]:

$$W_\psi(a,b)f = \frac{1}{\sqrt{C_\Psi}} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{|a|}} \Psi\left(\frac{b-x}{a}\right) f(x)\, dx, \tag{66}$$

где функция $\Psi$ называется вейвлетом; $b$ — смещение; $a$ — масштаб или шкала. Нормирующий коэффициент равен

$$C_\Psi = 2\pi \int\limits_{-\infty}^{\infty} \frac{|\tilde{\Psi}(\omega)|^2}{|\omega|}\, d\omega \;<\; \infty, \tag{67}$$

где $\tilde{\Psi}(\omega)$ — фурье-образ вейвлета $\Psi(x)$. Условие $C_\psi < \infty$ является условием существования вейвлета $\Psi$. Оно выполняется, в частности, если равны нулю первые $n-1$ моментов:

$$\int\limits_{-\infty}^{\infty} |x|^m \Psi(x) dx = 0, \quad 0 \le m < n. \tag{68}$$

Очевидно, что при $\Psi_{1/\omega,0}(x) = \mathrm{e}^{-i\omega x}$ мы имеем обычное преобразование Фурье. Обратное вейвлет-преобразование записывается в виде

$$f(x) = \frac{1}{\sqrt{C_\Psi}} \int\int \frac{1}{\sqrt{a}} \Psi\left(\frac{x-b}{a}\right) [W_\Psi(a,b)f]\, \frac{da\, db}{a^2}. \tag{69}$$

Свобода в выборе базисных функций $\Psi_{a,b}(x)$ позволила ввести многие типы вейвлетов, обычно называемые по имени предложивших их исследователей: вейвлеты Хаара, Добеши, Маллата, Мейера и т. д. [117]. В последние годы появилось также семейство быстровычисляемых вейвлетов второго поколения, порождаемых так называемой лифтинг-схемой [118]. В этой связи в работе [122] был проведен сравнительный анализ вычислительных и точностных свойств вейвлетов первого и второго поколений, чтобы помочь экспериментаторам в них ориентироваться.

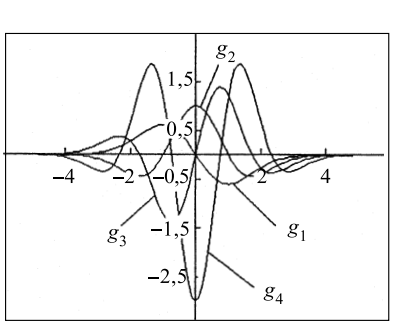
Часто в качестве вейвлетов используют производные функции Гаусса

$$\Psi(x) \equiv g_n(x) = (-1)^{n+1} \frac{d^n}{dx^n} \mathrm{e}^{-x^2/2}, \quad n > 0, \tag{70}$$

с нормирующим коэффициентом $C_{g_n} = 2\pi(n-1)!$, называемые поэтому гауссовыми вейвлетами (ГВ) или вейвлетами с нулевыми моментами. Первые два гауссова вейвлета хорошо известны [117]:

$$g_1(x) = -x\, \mathrm{e}^{-x^2/2}, \quad g_2(x) = (1-x^2)\, \mathrm{e}^{-x^2/2}. \tag{71}$$

Второй из них называется «мексиканской шляпой» (МНАТ).



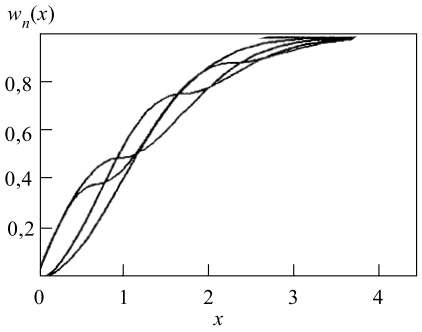




Рис. 26. Гауссовы вейвлеты от первого до четвертого порядка

Рис. 27. Относительные площади гауссовых вейвлетов

Однако, как выяснилось в работе [119], в задаче разрешения близких сигналов потребовались гауссовы вейвлеты и более высокого порядка, а именно:

$$g_3(x) = (3x - x^3)\,\mathrm{e}^{-x^2/2},$$

$$g_4(x) = (6x^2 - x^4 - 3)\,\mathrm{e}^{-x^2/2}.$$

Вид гауссовых вейвлетов от первого до четвертого порядка показан на рис. 26. Среди полезных свойств гауссовых вейвлетов отметим два, относящихся к их производным и интегралам:

$$\frac{d\,g_n(x)}{dx} = -g_{n+1}(x), \qquad \int\limits_{x_1}^{x_2} g_n(x) = g_{n-1}(x_1) - g_{n-1}(x_2). \tag{72}$$

Весьма важное свойство ГВ было найдено в [119], оно состоит в сохранении относительной площади ГВ (см. рис. 27), которая определяется как

$$w(x) = \frac{\int\limits_0^x |g_n(x)|\,dx}{\int\limits_0^\infty |g_n(x)|\,dx}. \tag{73}$$

Это свойство гауссовых вейвлетов сохранять относительную площадь при переходе к функциям более высокого порядка не зависит ни от смещения вейвлетов (что вполне очевидно), ни от их растяжения. С его помощью удалось выбрать оптимальные масштабы для ГВ разных порядков в задаче

оценки параметров сигналов колоколообразной формы. Их аппроксимация с помощью гауссиана

$$g(x; A, x_0, \sigma) = A \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right) \tag{74}$$

позволяет воспользоваться тем замечательным свойством, что гауссово вейвлет-преобразование от (74) выглядит так же, как и соответствующий вейвлет:

$$W_{g_n}(a,b)g = \frac{A\sigma a^{n+1/2}}{\sqrt{(n-1)!}s^{n+1}} g_n\left(\frac{b-x_0}{s}\right),$$

где обозначено $s = \sqrt{a^2+\sigma^2}$. Исключая $\exp\left(-\dfrac{(b-x_0^2)}{2(a^2+\sigma^2)}\right)$, с помощью отношения вейвлетов разных порядков авторы работы [119] получили явные выражения для оценки параметров сигнала гауссовой формы:

$$x_0 = b \pm \sqrt{(a^2+\sigma^2)\left[3 + \frac{\sqrt{2}\,(a^2+\sigma^2)}{a^2}\frac{W_{g_3}(a,b)g}{W_{g_1}(a,b)g}\right]}. \tag{75}$$

Верный знак в (75) можно получить, вычисляя коэффициенты нечетных вейвлетов в числителе и знаменателе в точках, лежащих достаточно далеко от предполагаемого положения сигнала. Аналогично через отношение четных вейвлетов, вычисленных в точках, лежащих как можно ближе к предполагаемому значению $x_0$, можно оценить и параметр полуширины сигнала $\sigma$:

$$\sigma^2 = -a^2\left(1 + \sqrt{\frac{3}{2}}\frac{W_{g_2}(a,x_0)g}{W_{g_4}(a,x_0)g}\right). \tag{76}$$

В [119] описан также метод оценки параметров дублета из двух близких сигналов гауссовой формы. Сравнение точностей этой оценки с аналогичной оценкой, полученной с помощью преобразования Фурье, показало преимущество вейвлет-оценок [120], приближающихся по точкам к предельным возможным значениям, вычисленным в [121]. Полезно проведенное в [119] исследование влияния на результаты этих оценок таких реально присутствующих факторов, как наличие случайного разброса измерений, грубость оцифровки сигналов и наличие «мертвых» каналов при оцифровке.

Помимо других многочисленных применений вейвлет-преобразований в ФВЭ для представления и анализа экспериментальных данных (см., например, обзор Н. М. Астафьевой [117] и статьи И. М. Дремина [98,123]), вейвлет-анализ особенно активно используется в последние годы для изучения взаимодействий ядер высоких энергий.

В работах [123] вейвлет-анализ использовался для распознавания образов в Pb–Pb-взаимодействиях при энергии 158 ГэВ/нуклон. В угловых распределениях вторичных частиц были найдены структуры, которые можно интерпретировать как черенковское излучение глюонов. В [123] угловое распределение частиц в событии представлялось в виде

$$\frac{d^2n}{d\varphi d\theta} = \frac{1}{N}\sum_{i=1}^{N}\delta(\varphi-\varphi_i)\delta(\theta-\theta_i), \tag{77}$$

где $\theta$ — полярный угол, отсчитываемый от направления импульса налетающего ядра; $\varphi$ — азимутальный угол; $N$ — число частиц в событии; $\varphi_i$ и $\theta_i$ — углы испускания $i$-го адрона; $\delta$ — функция Дирака.

Интегрируя (77) по $\theta$ или $\varphi$, получаем распределения по азимутальному или полярному углу соответственно:

$$\frac{dn}{d\varphi} = \frac{1}{N}\sum_{i=1}^{N}\delta(\varphi-\varphi_i), \tag{78}$$

$$\frac{dn}{d\theta} = \frac{1}{N}\sum_{i=1}^{N}\delta(\theta-\theta_i). \tag{79}$$

В физике высоких энергий часто используется вместо $\theta$ псевдобыстрота $\eta = -\ln(\operatorname{tg}(\theta/2))$. Распределение по $\eta$ будем задавать в виде

$$\frac{dn}{d\eta} = \frac{1}{N}\sum_{i=1}^{N}\delta(\eta-\eta_i). \tag{80}$$

Вейвлет-преобразование функций типа (78)–(80) имеет простой вид

$$W_\Psi(a,b) = \frac{1}{N}\sum_{i=1}^{N} a^{-1/2}\Psi\left(\frac{x-b}{a}\right), \tag{81}$$

то есть каждой частице сопоставляется функция двух переменных. Вейвлет-образ события дается суммой этих функций.

В качестве примера на рис. 28 представлены вейвлет-образы (спектры) события с 6 частицами, имеющими псевдобыстроты $\eta_1 = 1$, $\eta_2 = 2,75$, $\eta_3 = 3,25$, $\eta_4 = 5$, $\eta_5 = 6$, $\eta_6 = 7$. По горизонтальной оси отложено смещение, по вертикальной — масштаб. Значения вейвлет-коэффициентов представлены различными оттенками серого цвета. Белый цвет соответствует максимальному значению коэффициента, черный — минимальному. Даны также плотности энергии $W_\Psi(a,b)^2$ и скалограммы $E_W(a) = \int W_\Psi(a,b)^2 db$.

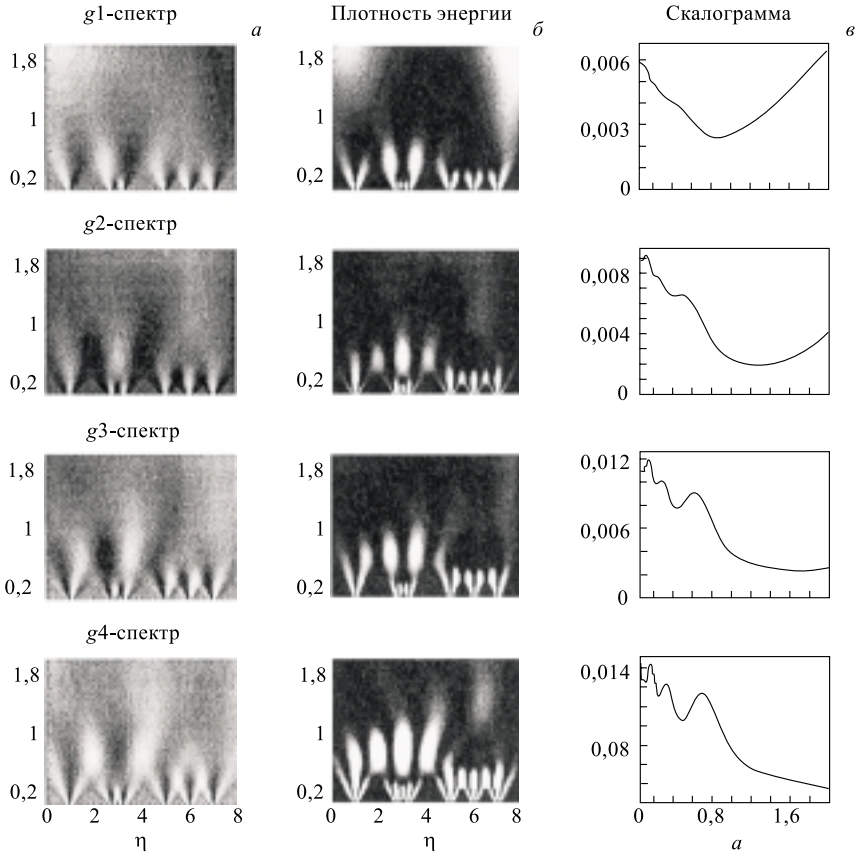



Рис. 28. Вейвлет-спектры (*а*), плотности энергии (*б*) и скалограммы тестового примера (*в*)

Вейвлет $g1$ (см. рис. 26) имеет максимум при отрицательных и минимум при положительных значениях своего аргумента. Положения этих точек экстремума на разных масштабах четко видны в спектре плотности энергии. Вейвлет $g2$ имеет 3 точки экстремума. Все они представлены в спектре энергии и т. д. Скалограммы дают обобщенное представление о точках экстремумов.

Как видно из рис. 28, разные вейвлеты дают разные представления частиц. Пожалуй, наиболее «привычные» получаются при использовании четных вейвлетов. Нечетные вейвлеты, возможно, могут быть полезны при автоматизированной обработке данных.

Вейвлет $g1$, на первый взгляд, позволяет локализовать области неоднородностей в распределении частиц — положения локальных максимумов плот-

ности энергии на шкале смещений ($b = 2,5$ и $b = 3,5$) выделяют группу частиц 2 и 3, а их положение на шкале масштабов — размеры группы. Однако следующие локальные максимумы располагаются при $a = 4$ и $b \sim 0$, $b \sim 9$. Какого-либо выделения группы частиц 4, 5, 6 не происходит. В то же время в соответствующей скалограмме имеются особенности, по-видимому, связанные с характеристиками групп частиц.

Более интересен вейвлет $g2$. Положения локальных максимумов ($a \simeq 0,5$ и $b = 3$, $a \simeq 1$ и $b = 6$) на шкале смещений связаны с положениями центров групп частиц 2, 3 и 4, 5, 6. Соответствующие масштабы дают представление о ширине группы на шкале псевдобыстрот. Положения локальных минимумов дают центры псевдобыстротных щелей между группами. В скалограмме имеются точки экстремума (минимумов) при $b \simeq 0,5$ и 1,2. Поэтому скалограммы можно использовать для быстрого выделения групп частиц.

Как видно из рис. 28, в вейвлет-спектре $g2$ при масштабе меньше 0,3 все частицы различимы. При масштабе больше 0,5 частицы 2 и 3 неразличимы. При $a > 1$ неразличимы частицы 4, 5, 6. При $a > 2$ можно было ожидать «слияния» частиц 1, 2 и 3. Однако этого не происходит из-за малого вклада частицы 1 в вейвлет-коэффициенты при больших масштабах. Таким образом, вейвлет-преобразование $g2$ позволяет группировать частицы. Аналогичными свойствами обладает вейвлет $g4$. К сожалению, характеристики соответствующих скалограмм не так четко скоррелированы с характеристиками частиц. Поэтому использовался в основном вейвлет $g2$.

В работе [124] этот подход использовался для изучения экспериментального материала, полученного при облучении на ускорителе SPS (CERN) ядрами кислорода и серы с импульсами 200 и 60 ГэВ/с/нуклон стопок ядерной фотоэмульсии НИКФИ БР-2. Исследовались 884 события $\mathrm{S} + \mathrm{Em}$- и 504 события $\mathrm{O} + \mathrm{Em}$-взаимодействий при 200 ГэВ/с/нуклон и 504 события $\mathrm{O} + \mathrm{Em}$-взаимодействий при 60 ГэВ/с/нуклон. В каждом найденном при просмотре событии для всех заряженных частиц были измерены пространственные ($\theta$) и азимутальные ($\varphi$) углы.

Рассматривались ливневые, или $s$-частицы, т. е. однозарядные частицы со скоростью $\beta = v/c \geq 0,7$, состоящие, в основном, из частиц, рожденных во взаимодействии, и однозарядных фрагментов снаряда. Как видно из рис. 29, распределение для $\mathrm{O} + \mathrm{Em}$-взаимодействий содержит две подструктуры в центральной области, проявляющиеся при масштабе 0,4. Не исключено, что структура может быть связана с особенностями ядра кислорода (ядра кислорода и углерода считаются сильно кластеризованными ядрами).

Характеристики событий, содержащих три группы частиц, даны на рис. 30. Можно заметить большое сходство с тестовым примером на рис. 28 и также проследить определенное расположение центров групп.

Следут отметить, что визуальная классификация событий страдает субъективизмом. Одно и то же событие разные наблюдатели могут отнести к собы-

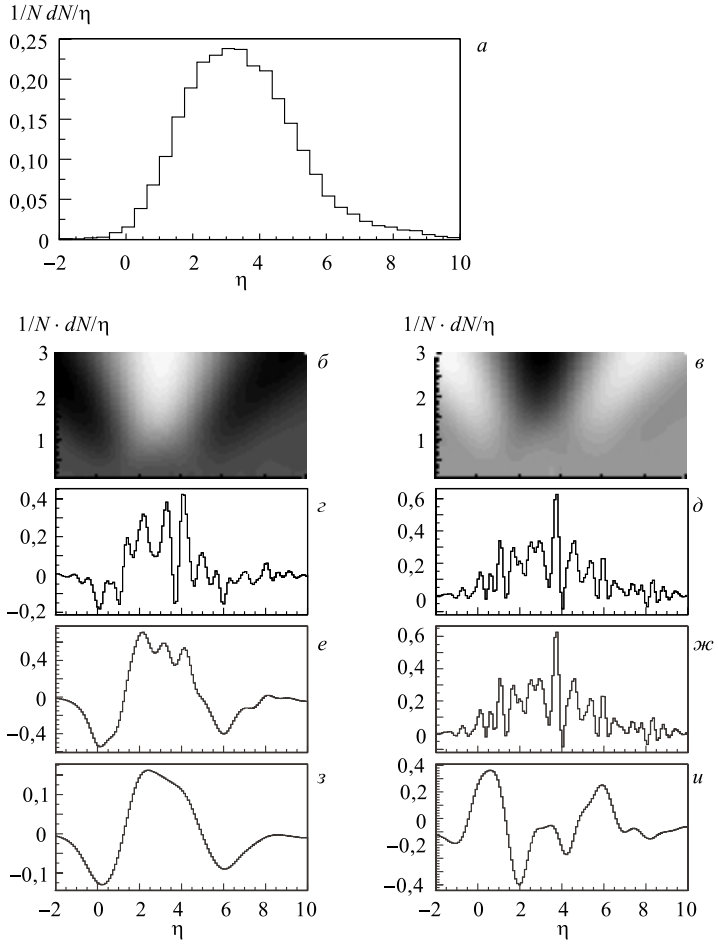



Рис. 29. Характеристики $s$-частиц в $\mathrm{O}+\mathrm{Em}$-взаимодействиях при 200 ГэВ/нуклон для $g2$- (*б*, *г*, *е*, *з*) и $g4$-спектров (*в*, *д*, *ж*, *и*) при $a = 0,2$ (*г*, *д*), $a = 0,4$ (*е*, *ж*) и $a = 0,6$ (*з*, *и*)

тиям с одной или двумя группами частиц. К счастью, скалограммы событий обладают особенностями, ассоциируемыми с картиной вейвлет-коэффициентов. Как указано в работе [124], скалограммы событий с одной группой частиц не имеют особых точек в интервале масштабов $0,5 \div 1,5$. Скалограммы событий с двумя или тремя группами частиц имеют особенности (точки экстремумов или перегибов) в указанном интервале масштабов. Положения особых точек скалограмм определяют характерные масштабы события. Число

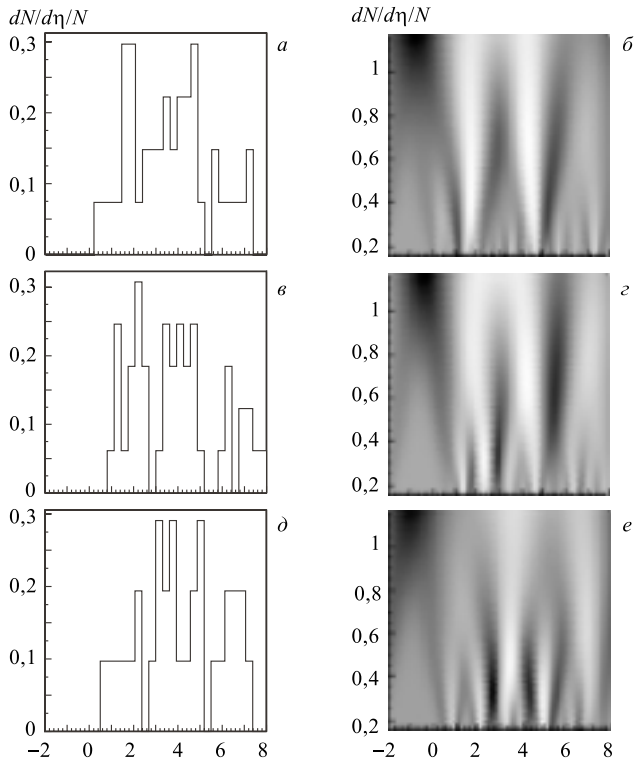



Рис. 30. Характеристики отдельных событий $\mathrm{S}+\mathrm{Em}$-взаимодействий при энергии 200 ГэВ/нуклон с тремя группами частиц: *а, б*) событие 42, $N_s = 49$; *в, г*) событие 189, $N_s = 77$; *д, е*) событие 287, $N_s = 33$

максимумов в спектре $g2$ коэффициентов на этих масштабах коррелирует с числом групп частиц, хотя о природе этих корреляций в работе не говорится.

Таким образом, вейвлет-преобразование псевдобыстротных распределений позволяет группировать частицы в события, причем для событий, содержащих такие группы, скалограммы имеют особые точки — точки экстремумов или перегибов. Их положения определяют характерные масштабы события. Особые точки вейвлет-коэффициентов при характерных масштабах коррелируют с характеристиками групп частиц.

Указанные закономерности позволяют выделять события с доминирующим рождением частиц в области фрагментации ядра-мишени, в центральной области или в области фрагментации ядра-снаряда. При энергии 200 ГэВ/нуклон происходит четкое разделение областей фрагментации. При энергии

60 ГэВ/нуклон наблюдается сильное перекрытие области фрагментации ядра-мишени и центральной области.

Несмотря на то, что представление о трех областях концентрации рожденных частиц уже обосновалось в феноменологической картине ядро-ядерных взаимодействий, до настоящего времени не было предложено ни одного инструментального алгоритмического метода выделения областей, пока применение вейвлет-анализа не позволило это сделать.

Согласно оценкам [124], в $\mathrm{S}+\mathrm{Em}$-взаимодействиях при энергии 200 ГэВ/нуклон в 42 % событий имеется одна группа частиц, в 29 % — две группы частиц, в 10 % событий — три и большее число групп. Доля исследованных событий ($N_s \geq 10$) составляет 82 % от полного числа событий.

С помощью специального пакета программ WASP [90] эти исследования были продолжены [91]. Обнаруженные нерегулярности в распределениях узких групп вторичных частиц по псевдобыстротам можно интерпретировать как наличие преимущественных углов испускания групп частиц.

## ЗАКЛЮЧЕНИЕ

В обзоре проведено обсуждение трех основных методов обработки экспериментальных данных, активно используемых в последние годы в Объединенном институте ядерных исследований: робастных методов математической статистики, искусственных нейронных сетей и клеточных автоматов и вейвлет-анализа. Обзор сделан, главным образом, по работам, выполненным с участием сотрудников Лаборатории информационных технологий ОИЯИ, в том числе и в рамках международных коллабораций с крупными физическими центрами: CERN, DESY, BNL и др. Авторы постарались достаточно подробно осветить основные понятия обсуждаемых методов и привести наиболее полезные и перспективные примеры их применения.

В частности, многочисленные примеры успешных применений робастных подходов в самых разных экспериментальных условиях, равно как и подчеркнутое в конце разд. 3 замечательное совпадение робастного подхода и мощных методов эластичного трекинга, говорят об их дальнейшей перспективности.

В разделе о нейронных сетях отмечено, насколько интенсивнее используются в ФВЭ многослойные прямоточные сети в сравнении с полносвязными сетями. Возможно, причиной этого является то, что последние требуют больших временных затрат или параллельной аппаратной реализации. Тем не менее, как показывает один из примеров их применения, их потенциал может быть востребован в рамках гибридных систем. Стоит обратить внимание на такие оригинальные и обещающие виды ИНС, как «соревновательные» RBF-сети и управляемые ИНС.

Большие возможности сулит применение вейвлет-анализа. Пока еще мало освоены двумерные вейвлеты, применение которых может дать новое качество для анализа сложных неявных зависимостей.

Обсуждение вопросов, включенных в данный обзор, выполнено с разной степенью подробности в силу как научных интересов авторов, так и ограниченности объема.

## СПИСОК ЛИТЕРАТУРЫ


1. *Cramer H.* Mathematical Methods of Statistics. N. Y.: Princeton Univ. Press, 1946.
2. *Huber P.* Robust Statistics. N. Y.: Wiley, 1981.
3. *Чернов Н. И. и др.* Численный анализ робастных регрессионных методов. Сообщение ОИЯИ Р5-85-492. Дубна, 1985.
4. *Fletcher R. et al.* // Comp. J. 1971. V. 72, No. 3. P. 276.
5. *Mosteller F., Tukey W.* Data Analysis and Regression: a Second Course in Statistics. N. Y.: Addison–Wesley, 1977.
6. *Ososkov G.* Robust Regression for the Heavy Contaminated Sample // Proc. of the 2nd Intern. Tampere Conf. in Statistics, Tampere, Finland, 1987. P. 615–626.
7. *Kirkpatrick S. et al.* // Science. 1983. V. 22. P. 671.
8. *Chernov N., Kolganova E., Ososkov G.* Fitting of Circles Registered by a High Granularity Detector // Proc. of the 8th Joint EPS-APS Intern. Conf. on Physics Computing, PC'96 / Eds. P. Borcherds, M. Bubak, A. Maksymowicz. Krakov, 1996. P. 230–233.
9. *Baur R. et al.* // Nucl. Instr. Meth. A. 1994. V. 343. P. 87.
10. *Müller U. et al.* // Ibid. 1994. V. 343. P. 279–283.
11. *Bächler J. et al.* // Ibid. P. 273–275.
12. *Di Mauro A. et al.* // Ibid. P. 284–287.
13. *Seguinot J., Ypsilantis T.* // Ibid. P. 1–29.
14. *Agakishiev H. et al.* New Robust Fitting Algorithm for Vertex Reconstruction in the CERES Experiment // Nucl. Instr. Meth. A. 1997. V. 394. P. 225–231.
15. The COMPASS Collaboration. COMPASS Proposal. Preprint CERN/SPSLC 96-14. 1996.
16. *Agakichiev G. et al.* Cherenkov Ring Fitting Techniques for the CERES RICH Detectors // Nucl. Instr. Meth. A. 1996. V. 371. P. 243–247.
17. *James F., Roos M.* «Minuit» a System for Function Minimization and Analysis of the Parameter Errors and Correlations // Comp. Phys. Comm. 1975. V. 10. P. 343.
18. *Ososkov G.* Novel Approach in RICH Data Handling // Czech. J. Phys. 1999. V. 49/S2. P. 145–160.
19. *Kolganova E. A., Ososkov G. A.* Particle Identifying Algorithms for Raw RICH Detector Data // Czech. J. Phys. 1999. V. 49/S2. P. 169–172.
20. *Chernov N., Kolganova E., Ososkov G.* Robust Methods for the RICH Ring Recognition and Particle Identification // Nucl. Instr. Meth. A. 1999. V. 433. P. 274–278.
21. *Linka A. et al.* MCMC Solution to Circle Fitting Problem in Analysis of RICH Detector Data // Proc. of COMPSTAT'98, 1998. P. 383–388.

22. *Kunde G. et al.* STAR RICH Proposal. BNL. 1998.

23. *Golutvin I. et al.* Robust Estimations of Muon Track Segment Parameters in CMS Endcap Muon Chambers // Proc. of CHEP'98, Chicago, 1998.

24. *Ososkov G., Palichik V., Tikhonenko E.* Robust Technique with Sub-Optimal Weight Function for Track Fitting in CMS Muon Strip Chamber // Abst. of Europhysics Conf. on Comp. Phys. EPS, Granada, Spain, 1998. V. 22F. P. 323–324.

25. *Tikhonenko E. et al.* Robust Estimates of Track Parameters and Spatial Resolution CMS Muon Chambers // Comp. Phys. Comm. 2000. V. 126/1-2. P. 72–76.

26. *Baginyan S., Ososkov G.* Finding Tracks Detected by a Drift Tube System // Comp. Phys. Comm. 1998. V. 108, No. 1. P. 20–28.

27. ATLAS Technical Proposals. CERN/LHCC/94-43, LHCC/P2. Dec. 15, 1994.

28. *Grote H.* CERN-DD/81/01. 1981.

29. *Никитин В. А., Ососков Г. А.* Автоматизация измерений и обработки данных физического эксперимента. М.: Изд-во Моск. ун-та, 1986.

30. *Богданова Н., Гаджоков В., Ососков Г. А.* Математические проблемы калибровки автоматизированных измерительных систем для оптических трековых детекторов в физике высоких энергий // ЭЧАЯ. 1986. Т. 17, вып. 5. С. 982–1029.

31. *Hulsbergen W.* OTR Calibration for Run 14577. 27/11/2000.

32. *Mankel R.* A «Canonical» Procedure to Fix External Degrees of Freedom in the Internal Alignment of a Tracking System. HERA-B Note 99-087. 1999.

33. *Alcaraz J., Josa M. I., Pinto J. C.* Global Alignment of the Silicon Microvertex Detector. L3 Internal Note. 1724. Feb. 23, 1995.

34. ALEPH Collaboration. Alignment of the ALEPH Tracking Devices. CERN-PPE, 92-90. May 29, 1992.

35. *Caccia M., Stocchi A.* The Delphi Vertex Detector Alignment a Pedagogical Statistical Exercise. INFN/AE, 90/16. Nov. 23, 1990.

36. *Barannikova O. et al.* SVT Internal Alignment Package: Part A. STAR Note. BNL. 1997.

37. *Agakishiev H. et al.* Alignmet of Detectors at CERES-NA45. JINR Commun. E10-98-277. Dubna, 1998.

38. *Barannikova O. et al.* SVT Internal Alignment Package. STAR Note 0356. BNL. 1998.

39. *Barannikova O. et al.* Specifications of SVT Global Alignment Package. STAR Note 0364. BNL. 1998.

40. *Кисель И. В., Нескоромный В. Н., Ососков Г. А.* Применение нейронных сетей в экспериментальной физике // ЭЧАЯ. 1993. Т. 24, вып. 6. С. 1551–1595.

41. *Peterson C. et al.* JETNET 3.0: A Versatile Artificial Neural Network Package. LU TP 93-29. 1993; CERN-TH 7135/94.

42. *Lindsey C., Lindblad Th.* Review of Hardware Neural Networks: a User's Perspective // Intern. J. of Neural Syst. 1995. V. 6. (Suppl.).

43. Nucl. Instr. Meth. A. 1997. V. 389.

44. *Rumelhart D. et al.* Learning Internal Representation by Error Propagation // Explorations in the Microstructure of Cognition: Parallel Distributed Proc. Cambridge, 1986. V. 1.

45. *Bevington P.* Data Reduction and Error Analysis for the Physical Sciences. N. Y.: McGraw-Hill, 1969.

46. *Marquard D.* An Algorithm for Least Squares Estimations of Non-Linear Parameters // SIAM J. Appl. Math. 1903. V. 11. P. 431–441.

47. *Иванов В. В. и др.* Возможность применения нейронной сети для отбора событий с $B$-мезонами в калориметрическом триггере // Матем. моделир. 1992. Т. 4, № 8. С. 94–106.

48. *Ivanov V. et al.* An Algorithm for Identifying Events in the Experiment DISTO // Крат. сообщ. ОИЯИ. 1995. № 2[70].

49. *Puzynin I. et al.* A Newton-Type Algorithm for Multilayer Perceptron Training // Proc. of «New Computing Techniques in Physics Research III», Oberammergau, Germany. Singapore, 1993. P. 269–274.

50. *Palichik V. et al.* // Ibid. P. 347–352.

51. *Ермаков В. В., Калиткин Н. Н.* // ЖВМ и МФ. 1981. Т. 21, вып. 2. С. 491.

52. *Пузынин И. В. и др.* Алгоритмы отбора событий в вершинном детекторе спектрометра МЧС. Сообщение ОИЯИ Р10-94-300. Дубна, 1994.

53. *Бонюшкина А. Ю. и др.* Об одном выборе входных данных для многослойного персептрона. Сообщение ОИЯИ Р10-94-363. Дубна, 1994.

54. *Зрелов П. В., Иванов В. В.* Сообщение ОИЯИ Р10-92-461. Дубна, 1992.

55. *Лоули Д., Максвелл А.* Факторный анализ как статистический метод. М.: Мир, 1967.

56. Вероятность и математическая статистика // Энциклопедия. М., 1999.

57. *Hansroul M. et al.* JINR Commun. D10-73-10. Dubna, 1973.

58. *Ососков Г. А., Поспелов А. С.* Сообщение ОИЯИ Р10-91-444. Дубна, 1991; Р10-92-317. Дубна, 1992.

59. *Карманова И. В. и др.* Применение нейронных сетей для дифференциальной диагностики тяжести течения пневмонии // Тр. конф. «Физика и радиоэлектроника в медицине и экологии (ФРЭМЭ-2000)». Владимир, 2000.

60. *Ososkov G., Stadnik A.* Neural Network Application for the Face Recognition Systems. JINR Preprint E11-2000-269. Dubna, 2000.

61. *Ososkov G., Stadnik A.* Face Recognition by a New Type of Neural Networks // Advances in Neural Networks and Applications / Ed. N. Mastorakis. WSES Press, 2001. P. 304–308.

62. *Haykin S.* Neural Networks: a Comprehensive Foundation. N. Y., 1994.

63. http://www.vsma.ac.ru/ pharm/library/books/textbook/stneurnet.html

64. *Chellappa R. et al.* Human and Machine Recognition of Faces: a Survey // Proc. of IEEE. 1995. V. 83, No. 5. P. 704–740.

65. *Kondo T., Yan H.* Automatic Human Face Detection and Recognition under Non-Uniform Illumination // Pattern Recognition. 1999. V. 32(10). P. 1707–1718.

66. *Ranganath S., Arun K.* Face Recognition Using Transform Features and Neural Networks // Pattern Recognition. 1997. V. 30(10). P. 1615–1622.

67. *Li-Fen Chen et al.* A New LDA-Based Face Recognition System Which Can Solve the Small Sample Size Problem // Pattern Recognition. 2000. V. 33(10). P. 1713–1726.

68. ftp://ftp.uk.research.att.com:pub/data/att_faces.tar.Z

69. *Уоссермен Ф.* Нейрокомпьютерная техника: Теория и практика: Пер. с англ. / Под ред. А. И. Галушкина. М.: Мир, 1992. 238 с.

70. *Puzynin I. et al.* Dynamics of Realistic Neural Network with Time-Dependent External Signal // Comp. Math. Appl. 1997. V. 34, No. 7/8. P. 667–675.

71. *Кропотов Ю. Д., Пахомов С. В.* Физиология человека. 1984. Т. 10. С. 813.

72. *Hopfield J.* Learning Algorithms and Probability Distributions in Feedforward and Feedback Networks // Proc. Nat. Acad. Sci. USA. 1987. V. 84. P. 8429.

73. *Hopfield J.* Neural Networks and Physical Systems with Emergent Collective Computational Abilities // Proc. Nat. Acad. Sci. USA. 1982. V. 79. P. 2554.

74. *Hopfield J.* Neurons with Graded Responces Have Collective Computational Properties Like Those of Two-State Neurons // Proc. Nat. Acad. Sci. USA. 1984. V. 81. P. 3088.

75. *Peterson C.* Track Finding with Neural Networks // Nucl. Instr. Meth. A. 1986. V. 279. P. 537.

76. *Denby B.* Neural Networks and Cellular Automata in Experimental High Energy Physics // Comp. Phys. Comm. 1988. V. 49. P. 429.

77. *Peterson C., Söderberg B.* A New Method for Mapping Optimization Problems onto Neural Networks // Intern. J. of Neural Syst. 1989. V. 1. P. 3.

78. *Baginyan S. et al.* Tracking by Modified Rotor Model of Neural Network // Comp. Phys. Comm. 1994. V. 79. P. 165.

79. *Ососков Г. А. и др.* Использование нейронных сетей для улучшения интерпретации эксперимента EXCHARM // Матем. моделир. 1999. Т. 11, вып. 10. С. 116–126.

80. *Ososkov G.* Robust Tracking by Cellular Automata and Neural Network with Non-Local Weights // Applications and Science of Artificial Neural Networks / Eds. S. K. Rogers, D. W. Ruck. Proc. SPIE 2492. 1995. P. 1180–1192.

81. *Ososkov G., Zaznobina E.* Tracking by Cellular Automata and Neural Networks // Intern. J. of Neural Syst. 1995. V. 6 (Suppl.). P. 269–274;
*Зазнобина Е. Г., Ососков Г. А.* Сообщение ОИЯИ Р10-93-404. Дубна, 1993.

82. *Grozov V. P. et al.* Automatic Processing of Ionograms on the Basis of the Artificial Neural Network Method // Proc. of Intern. Symp. on Radio Propagation (ISRP'97), Qingdao, China, 1997. P. 514–517.

83. *Грозов В. П., Носов В. Е., Ососков Г. А.* Вопросы обработки изображений применительно к задачам автоматической обработки ионограмм // Тез. докл. IV симпоз. «Оптика атмосферы и океана», Томск, 1997. С. 79–80.

84. *Galkin I. et al.* Feedback Neural Networks for ARTIST Ionogram Processing // Radio Science. 1996. V. 31, No. 5. P. 1119–1128.

85. *Lindsey C. et al.* Experience with the IBM ZISC036 Neural Network Chip // Proc. of IV Intern. Workshop on Software Engineering and Artificial Intelligence and Expert Systems for High Energy and Nuclear Physics, Pisa, Italy, April 3–8, 1995 / Ed. by B. Denby, D. Perret-Gallix. Singapore, 1995. P. 371–376.

86. *Ososkov G. A., Tikhonenko E. A.* New Random Number Generator on the Base of 2D-Cellular Automaton // Ibid. P. 635–640.

87. *Baginyan S., Ososkov G.* Controlled Neural Network Application in Track-Match Problem // ibid. P. 731–736; Сообщение ОИЯИ Е10-93-415. Дубна, 1993.

88. *Bonushkina A. et al.* Input Data for a Multilayer Perceptron in the Form of Variational Series // ibid. P. 751–756.

89. *Bonushkina A. et al.* Multivariate Data Anylysis Based on the $\omega_n^k$-Criteria and Multilayer Perceptron // Comp. Math. Appl. 1996. V. 34, No. 7/8. P. 677.

90. *Altaisky M. et al.* WASP (Wavelet Analysis of Secondary Particles Angular Distributions Package). Version 1.2. Long Write Up and User's Guide. JINR Commun. P2-2001-205. Dubna, 2001.

91. *Ужинский В. В. и др.* Вейвлет-анализ угловых распределений вторичных частиц в ядро-ядерных взаимодействиях при высоких энергиях. Нерегулярности псевдобыстротных распределений частиц. Сообщение ОИЯИ Р2-2001-288. Дубна, 2001.

92. *Gyulassy M., Harlander M.* Elastic Tracking and Neural Networks Algorithms for Complex Pattern Recognition // Comp. Phys. Comm. 1991. V. 66. P. 31–46.

93. *Durbin R., Willshaw D.* // Nature. 1987. V. 326. P. 689.

94. *Kisel I. et al.* Elastic Net for Broken Multiple Scattered Tracks // Comp. Phys. Comm. 1996. V. 98. P. 45–51.

95. *Lasiuk B. et al.* Development of an Elastic Tracking Package // Proc. of CHEP'98, Chicago, 1998.

96. *Ohlsson M., Peterson C., Yuille A.* Track Finding with Deformable Templates — the Elastic Arms Approach // Comp. Phys. Comm. 1992. V. 71. P. 77.

97. *Muresan L. et al.* Deformable Templates for Circle Recognition // JINR Rapid Comm. 1997. No. 1[81].

98. *Дремин И. М.* // УФН. 2000. Т. 170. С. 1235.

99. *Glazov A. et al.* Filtering Tracks in Discrete Detectors Using a Cellular Automaton // Nucl. Instr. Meth. A. 1993. V. 329. P. 262–268.

100. *Barannikova O., Ososkov G., Panebratsev Yu.* Investigation of the Fast Algorithm for Track and Vertex Finding Using New Computational Model with the Real STAR-SVT Geometry. JINR Commun. E10-98-278. Dubna, 1998.

101. *Toffoli T., Margolius N.* Cellular Automata Machines: a New Environment for Modelling. Cambridge, 1987.

102. *Wolfram S.* Theory and Applications of Cellular Automata / Ed. by S. Wolfram. Singapore: World Scientific, 1986.

103. *Gardner M.* // Scientific American. 1979. V. 223(4). P. 120.

104. *Ivanov V. et al.* On a Possible Second-Level Trigger for the Experiment DISTO // Nuovo Cim. A. 1996. V. 109, No. 3. P. 327–339.

105. *Kisel I. et al.* Application of CA and NN for Event Recognition in Experiments DISTO and STREAMER // Nucl. Instr. Meth. A. 1997. V. 389. P. 208–209.

106. *Kisel I. et al.* Elastic Neural Net for Track and Vertex Search // Ibid. P. 167–168.

107. *Kisel I. et al.* Cellular Automaton and Elastic Net for Event Reconstruction in the NEMO-2 Experiment // Nucl. Instr. Meth. A. 1997. V. 387. P. 433–442.

108. *Kisel I., Masciocchi S.* CATS — a Cellular Automaton for Tracking in Silicon for the HERA-B Vertex Detector. HERA-B Note 99-242. 1999.

109. *Ососков Г. А., Тихоненко Е. А.* Новый генератор случайных чисел на базе двумерного клеточного автомата // Матем. моделир. 1996. Т. 8, № 12. С. 77–84.

110. *Knuth D.* Seminumerical Algorithms. Addison-Wesley, 1981. V. 2.

111. *Ososkov G., Tikhonenko E., Zadorozhnyi A.* On Experience of Testing New Random Number Generator on the Base of $2D$-Cellular Automaton // Proc. of the 10th Summer School on Comp. Technique in Physics, Skalsky Dvur, Czech Rep. Sept. 5–14, 1995.

112. *James F.* A Review of Pseudorandom Number Generators // Comp. Phys. Comm. 1990. V. 60. P. 329–344.

113. *Gyulassy M., Harlander M.* High Resolution Multiparticle Tracking without Preprocessing Via Elastic Tracking // Nucl. Instr. Meth. A. 1992. V. 316. P. 238–245.

114. *Chernov N., Ososkov G., Silin I.* Robust Fitting of Ellipses to Non-Complete and Contaminated Data // Czech. J. Phys. 2000. V. 50 (Suppl. S1). P. 347–354.

115. *Kurbatov V., Silin I.* // Nucl. Instr. Meth. A. 1994. V. 345. P. 346–350.

116. Компьютерра. 1998. № 8(236).

117. Proc. of Symp. in Appl. Math. / Ed. by I. Daubechies. 1993. V. 47;
*Астафьева Н. М.* // УФН. 1996. Т. 166. С. 1145.

118. *Sweldens W., Schroder P.* Building Your Own Wavelets at Home in «Wavelets in Computer Graphics». ACM SIGGRAPH Course Notes. 1996. P. 15–87.

119. *Ososkov G., Shitov A.* Gaussian Wavelet Features and Their Applications for Analysis of Discretized Signals // Comp. Phys. Comm. 2000. V. 126/1-2. P. 149–157; Сообщение ОИЯИ Р11-97-347. Дубна, 1997.

120. *Altaisky M.V. et al.* // Proc. of the Intern. Conf. of SPIE'96. 1996. V. 2847. P. 656–664.

121. *Kolganova E., Kosarev E., Ososkov G.* Superresolution Algorithms for Data Analysis of Discrete Detectors in Nuclear Physics // Nucl. Instr. Meth. A. 2000. V. 443/2-3. P. 464–477.

122. *Ososkov G., Shitov A., Stadnik A.* Comparative Study of Some of Wavelets of the First and the Second Generations. JINR Commun. E11-2001-38. Dubna, 2001.

123. *Astafyeva N. M., Dremin I. M., Kotelnikov K. A.* // Mod. Phys. Lett. A. 1997. V. 12. P. 1185;
*Dremin I. M. et al.* hep-ph/0007060.

124. *Ужинский В. В. и др.* Вейвлет-анализ угловых распределений вторичных частиц в ядро-ядерных взаимодействиях при высоких энергиях. Сообщение ОИЯИ Р1-2001-119. Дубна, 2001.