УДК 51-72:539.12.043

# О СВЯЗИ МЕЖДУ МЕТОДАМИ МОМЕНТОВ И МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

*Ф. В. Ткачев*

Институт ядерных исследований РАН, Москва

Простой и общий критерий построения обобщенных моментов позволяет приближаться по точности к теоретическому пределу Крамера–Рао при оценивании параметров, избегая трудностей метода максимального правдоподобия в случаях со сложной плотностью вероятности или с большими массивами событий.

A simple and general criterion is presented for a practical construction of generalized moments that allow one to approach the theoretical Cramer–Rao limit for parameter estimation while avoiding the complexity of the maximum likelihood method in the cases of complicated probability distributions and/or very large event samples.

Существует простая и фундаментальная связь между методами обобщенных моментов и максимального правдоподобия (см. ниже уравнение (20)). Эта связь уже была эффективно использована в решении задачи об оптимальном алгоритме определения адронных струй [1]. Однако ее эвристическое значение шире: получающиеся из нее практические рецепты (которые удобно назвать методом квазиоптимальных наблюдаемых) могли бы быть полезны во многих задачах физики высоких энергий. В литературе эта связь, по-видимому, не имела того обсуждения, которого она, очевидно, заслуживает; во всяком случае о ней нет упоминания ни в одной из более десятка проверенных автором книг (включая известное руководство [2] и систематическое математическое изложение [3]), ни, скажем, в обзоре методов, использовавшихся при открытии и в исследованиях топ-кварка, данном в [4]; заметим, что в подобных исследованиях ввиду их сложности задействованы практически все известные методы статистического оценивания параметров. Факт отсутствия следов использования обсуждаемой связи в экспериментальной практике (несмотря на почтенный возраст обоих методов) можно объяснить тем, что для ее выявления нужно принять функционально-аналитическую точку зрения и провести аналогию между задачей минимизации дисперсии как функционала обобщенных моментов и теорией экстремумов функций, определенных на конечномерных областях, что достаточно непривычно. Получающийся «метод квазиоптимальных наблюдаемых» составляет альтернативу методу максимального правдоподобия в задачах, где применение последнего непрактично или невозможно, скажем, из-за сложности или отсутствия теоретических выражений для плотности вероятности (как это обычно имеет место в физике высоких энергий, где теоретические описания процессов часто принимают форму генераторов событий) или из-за большого размера выборки событий.

Пусть $\mathbf{P}$ — случайная переменная, конкретные значения которой будем называть событиями. Их плотность вероятности обозначим $\pi(\mathbf{P})$. Предполагается, что плотность зависит от параметра $M$, который следует оценить из экспериментальной выборки событий $\{\mathbf{P}_i\}_i$. Метод обобщенных моментов заключается в выборе некоторой функции

$f\mathbf{P}$ на событиях (обобщенный момент) и в оценке $M$ фитированием теоретического среднего,

$$\langle f \rangle = \int d\mathbf{P}\, \pi(\mathbf{P}) f(\mathbf{P}), \tag{1}$$

с соответствующим экспериментальным значением

$$\langle f \rangle_{\exp} = \frac{1}{N} \sum_i f(\mathbf{P}_i). \tag{2}$$

Задача состоит в нахождении момента $f$, который позволил бы извлечь $M$ с наилучшей точностью из данной выборки.

В пределе больших $N$ флуктуации оценки $M$ связаны с флуктуациями значений среднего $\langle f \rangle$ формулой

$$\delta M = \left( \frac{\partial \langle f \rangle}{\partial M} \right)^{-1} \delta \langle f \rangle. \tag{3}$$

Производную следует применять только к плотности вероятностей:

$$\frac{\partial \langle f \rangle}{\partial M} = \int d\mathbf{P} f(\mathbf{P}) \frac{\partial \pi(\mathbf{P})}{\partial M}, \tag{4}$$

т. к. фитирование $M$ производится с фиксированным $f$. (Здесь сделано обычное предположение, что границы области изменения $\mathbf{P}$ не зависят от $M$.) Для малых флуктуаций $\partial \langle f \rangle = N^{-1/2} \sqrt{\operatorname{Var} f}$, где

$$\operatorname{Var} f = \int d\mathbf{P}\, \pi(\mathbf{P}) (f(\mathbf{P}) - \langle f \rangle)^2 \equiv \langle f^2 \rangle - \langle f \rangle^2, \tag{5}$$

В терминах дисперсии уравнение (3) принимает вид

$$\operatorname{Var} M[f] = \left( \frac{\partial \langle f \rangle}{\partial M} \right)^{-2} \operatorname{Var} f; \tag{6}$$

задача состоит в минимизации этого выражения подходящим выбором $f$.

В дальнейших рассуждениях (функциональные производные и т. п.) мы не будем стремиться к формальной строгости, т. к. это затемнило бы простую аналогию с изучением минимумов обычных функций, не принося практической пользы. Достаточно заметить, что диапазон применимости получающихся рецептов таков же, как и у метода максимального правдоподобия, и что наш вывод имеет близкие параллели в стандартных доказательствах неравенства Крамера–Рао в терминах гильбертовой нормы (см., например, [5]).

Необходимое условие минимума может быть записано в терминах функциональных производных:

$$\frac{\delta}{\delta f(\mathbf{P})} \operatorname{Var} M[f] = 0. \tag{7}$$

Подставим уравнение (6) в (7) и используем соотношения

$$\frac{\delta}{\delta f(\mathbf{P})}\langle f\rangle = \pi(\mathbf{P}),\ \frac{\delta}{\delta f(\mathbf{P})}\langle f^2\rangle = 2f(\mathbf{P})\pi(\mathbf{P}),\ \frac{\delta}{\delta f(\mathbf{P})}\frac{\partial\langle f\rangle}{\partial M} = \frac{\partial\pi(\mathbf{P})}{\partial M}. \tag{8}$$

Простые вычисления дают

$$f(\mathbf{P})\langle f\rangle + \mathrm{const}\frac{\partial\ln\pi(\mathbf{P})}{\partial M}, \tag{9}$$

где константа не зависит от $\mathbf{P}$. Эта константа роли не играет, т. к. $f$ определяется данным рассуждением только с точностью до постоянного множителя. Замечая, что

$$\int d\mathbf{P}\,\pi(\mathbf{P})\frac{\partial\ln\pi(\mathbf{P})}{\partial M} = \frac{\partial}{\partial M}\int d\mathbf{P}\,\pi(\mathbf{P}) \equiv \frac{\partial}{\partial M}1 = 0, \tag{10}$$

приходим к следующему общему семейству решений:

$$f(\mathbf{P})C_1\frac{\partial\ln\pi(\mathbf{P})}{\partial M} + C_2, \tag{11}$$

где $C_i$ не зависят от $\mathbf{P}$, но могут зависеть от $M$.

Для удобства, как обычно, имеем в виду следующий член семейства (11):

$$f_{\mathrm{opt}}(\mathbf{P}) = \frac{\partial\ln\pi(\mathbf{P})}{\partial M}. \tag{12}$$

Тогда уравнение (10) в сущности совпадает с

$$\langle f_{\mathrm{opt}}\rangle = 0. \tag{13}$$

В качестве простого примера рассмотрим распределение Брейта–Вигнера. Пусть $\mathbf{P}$ — случайные вещественные числа, распределенные в соответствии с

$$\pi(\mathbf{P}) \propto \frac{1}{(M-\mathbf{P})^2+\Gamma^2} \tag{14}$$

в некотором фиксированном интервале, включающем точку $\mathbf{P} = M$. Тогда оптимальным моментом для оценки $M$ будет

$$f_{M,\mathrm{opt}}(\mathbf{P}) = \frac{\partial}{\partial M}\ln\pi(\mathbf{P}) = -\frac{2(M-\mathbf{P})}{(M-\mathbf{P})^2+\Gamma^2}. \tag{15}$$

(Напомним, что аддитивные и мультипликативные не зависящие от $\mathbf{P}$ константы могут быть опущены в таких выражениях; см. уравнение (11).)

Интересно отметить, как $f_{M,\mathrm{opt}}$ подчеркивает вклады боковых частей пика — в точности там, где величина $\pi(\mathbf{P})$ наиболее чувствительна к вариациям $M$, причем учет вкладов от двух боковых частей с противоположным знаком максимизирует сигнал. При этом выражение (15) подавляет вклады от значительного числа событий в средней части пика (14), которые представляют собой, скорее, шум с точки зрения измерения $M$.

Уравнение (12) представляет собой перевод метода максимального правдоподобия (который, как известно, дает теоретически наилучшую оценку для $M$; см. неравенство Крамера–Рао [2, 3]) на язык обобщенных моментов. В самом деле, метод максимального правдоподобия предлагает оценивать $M$ значением, максимизирующим функцию правдоподобия

$$\sum_i \ln \pi(\mathbf{P}_i), \tag{16}$$

где суммирование ведется по всем событиям в выборке. Необходимое условие для максимума имеет вид

$$\frac{\partial}{\partial M}\left[\sum_i \ln \pi(\mathbf{P}_i)\right] = \sum_i \frac{\partial \ln \pi(\mathbf{P}_i)}{\partial M} \propto \langle f_{\rm opt}\rangle_{\rm exp} = 0. \tag{17}$$

Это согласуется с (12) благодаря (13). Удивительно, что в общедоступной литературе, во-видимому, отсутствует явная и общая формулировка метода максимального правдоподобия в терминах обобщенных моментов, хотя эквивалентные формулы можно обнаружить в примерах вывода конкретных оценок для параметров стандартных распределений, например, в [5].

Изучим, как малые отклонения от $f_{\rm opt}$ влияют на точность оценки $M$. Рассмотрим функционал $\mathrm{Var}\, M[f]$. Пусть $\phi$ — функция на событиях, причем $\langle \phi^2 \rangle < \infty$. Напишем функциональный аналог разложения Тейлора для $\mathrm{Var}\, M[f_{\rm opt}+\phi]$ по $\phi$, удерживая квадратичные члены:

$$\mathrm{Var}\, M[f_{\rm opt}+\phi] = \mathrm{Var}\, M[f_{\rm opt}] + \frac{1}{2}\int \left[\frac{\delta^2 \mathrm{Var}\, M[f]}{f(\mathbf{P})\,\delta(\mathbf{Q})}\right]_{f=f_{\rm opt}} \phi(\mathbf{P})\phi(\mathbf{Q})\, d\mathbf{P}\, d\mathbf{Q} + \dots . \tag{18}$$

Линейный по $\phi$ член не возникает, т. к. $f_{\rm opt}$ удовлетворяет (7). Чтобы вычислить квадратичный член в (18), достаточно использовать функциональные производные и соотношения типа (8) и

$$\frac{\delta}{\delta f(\mathbf{P})} f(\mathbf{Q}) = \delta(\mathbf{P},\mathbf{Q}), \quad \int \delta(\mathbf{P},\mathbf{Q})\phi(\mathbf{P}) d\mathbf{P} = \phi(\mathbf{Q}). \tag{19}$$

Прямой расчет дает главный технический результат данной работы:

$$\mathrm{Var}\, M[f_{\rm opt}+\phi] = \frac{1}{\langle f_{\rm opt}^2\rangle} + \frac{1}{\langle f_{\rm opt}^2\rangle^3}\left\{\langle f_{\rm opt}^2\rangle \times \langle \bar\phi^2\rangle - \langle f_{\rm opt}\times\bar\phi\rangle^2\right\} + \dots , \tag{20}$$

где $\bar\phi = \phi - \langle\phi\rangle$. Неотрицательность выражения в фигурных скобках следует из неравенства Буняковского–Коши–Шварца, так что мы действительно имеем дело с минимумом. (Заметим, что неравенство Буняковского–Коши–Шварца фигурирует и в стандартных строгих доказательствах теоремы Крамера–Рао.)

Первый член в правой части (20), $\langle f_{\rm opt}^2\rangle^{-1}$, дает абсолютный минимум дисперсии $M$, установленный неравенством Крамера–Рао. Последнее справедливо для всех $\phi$ и поэтому несколько сильнее, чем результат (20), полученный лишь для достаточно малых $\phi$. Однако уравнение дает простое явное описание отклонений от оптимальности и поэтому делает возможным практические предписания, данные ниже после уравнения (24).

Удобно говорить об *информативности* $I_f$ обобщенного момента $f$ по отношению к параметру $M$, определенной как

$$I_f = (\operatorname{Var} M[f])^{-1}. \tag{21}$$

Информативность оптимального момента $f_{\rm opt}$ равна

$$I_{\rm opt} = \langle f_{\rm opt}^2 \rangle, \tag{22}$$

что соответствует пределу Крамера–Рао. Разложение явно описывает отклонения от предела. Информативность тесно связана с информацией Фишера [2, 3], которая, однако, представляет собой свойство данных, тогда как информативность — свойство момента.

Тот факт, что решение является точкой квадратичного минимума, означает, что любой момент $f_{\rm quasi}$, достаточно близкий к (12), будет на практике так же хорош, как и оптимальное решение (назовем такие моменты квазиоптимальными). Количественная мера близости дается сравнением членов порядков $O(1)$ и $O(\phi^2)$ в правой части (20):

$$\frac{\langle f_{\rm opt}^2 \rangle \langle \bar\phi^2 \rangle - \langle f_{\rm opt} \bar\phi \rangle^2}{\langle f_{\rm opt}^2 \rangle^2} \ll 1, \tag{23}$$

где $\bar\phi = f_{\rm quasi} - \langle f_{\rm quasi} \rangle - f_{\rm opt}$.

Вычитаемый член в числителе неотрицателен, так что, опуская его, получим достаточное условие для уравнения, тем более, что $\langle f_{\rm opt} \bar\phi \rangle$ будет подавляться, если $f_{\rm quasi}$ осциллирует около $f_{\rm opt}$. Предполагая без потери общности, что $\langle f_{\rm quasi} \rangle = 0$, получаем следующий удобный достаточный критерий:

$$\langle [f_{\rm quasi} - f_{\rm opt}]^2 \rangle \ll \langle f_{\rm opt}^2 \rangle. \tag{24}$$

Принимая во внимание это и уравнение (20) и обозначая стандартное отклонение $\sigma$ для $M$ в оптимальном и квазиоптимальном случаях как $\sigma_{\rm opt}$ и $\sigma_{\rm quasi}$, соответственно, получим

$$\frac{\sigma_{\rm quasi}}{\sigma_{\rm opt}} \approx 1 + \frac{1}{2} \frac{\langle [f_{\rm quasi} - f_{\rm opt}]^2 \rangle}{\langle f_{\rm opt}^2 \rangle}. \tag{25}$$

Теперь метод **квазиоптимальных моментов** можно сформулировать следующим образом:

1) выполним (любым способом) приближенное построение обобщенного момента $f_{\rm quasi}$, руководствуясь формулой (12), так чтобы $f_{\rm quasi}$ был близок к $f_{\rm opt}$ в интегральном смысле уравнению (24);

2) определим $M$ фитированием $\langle f_{\rm quasi} \rangle$ и $\langle f_{\rm quasi} \rangle_{\rm exp}$;

3) оценим ошибку для $M$ посредством (6);

4) если $f_{\text{quasi}}$ достаточно чувствительно зависит от $M$, то можно использовать итеративную процедуру, начав с некоторого $M_0$, близкого к истинному (в задачах прецизионных измерений такое $M_0$ всегда можно указать; либо можно использовать альтернативный метод, скажем, $\chi_2$), и затем модифицировать $f_{\text{quasi}}$ после получения новой оценки для $M$ (это напоминает поиск экстремума в методе максимального правдоподобия).

Для практического построения квазиоптимальных моментов $f_{\text{quasi}}$ полезно переформулировать в терминах подынтегральных выражений:

$$\int d\mathbf{P}\pi(\mathbf{P})[f_{\text{quasi}}(\mathbf{P}) - f_{\text{opt}}(\mathbf{P})]^2 \ll \int d\mathbf{P}\pi(\mathbf{P}) f_{\text{opt}}^2(\mathbf{P}). \tag{26}$$

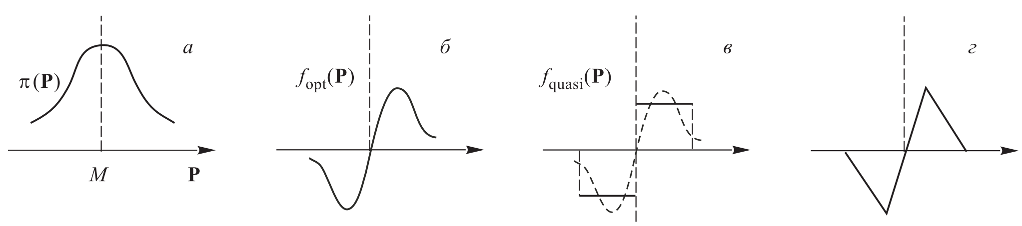
В качестве наводящего соображения следует минимизировать выражение в скобках в левой части (26):

$$[f_{\text{quasi}}(\mathbf{P}) - f_{\text{opt}}(\mathbf{P})]^2 \ll f_{\text{opt}}^2(\mathbf{P}). \tag{27}$$

Неравенство должно выполняться для «большинства» $\mathbf{P}$, т. е. следует учитывать величину $\pi(\mathbf{P})$: неравенство (27) можно ослабить в областях, дающих малый вклад в интеграл в левой части (26).

Рассмотрим в качестве примера уравнение (14). Предположим, что точное распределение вероятностей отличается от (14), скажем, слабой, но сложной зависимостью $\Gamma$ от $\mathbf{P}$ (что может иметь место в вычислениях по теории возмущений — типичная ситуация в физике высоких энергий [6]). Тогда правая часть (15) с постоянной $\gamma$ будет соответствовать обобщенному моменту, который всего лишь квазиоптимален, но отклонения от оптимальности могут быть практически пренебрежимы (в зависимости от того, насколько слабой является зависимость от $\mathbf{P}$). Тогда можно использовать момент, задаваемый простейшей формулой (15), практически без потери информативности.

Также можно заменить аналитическую кривую кусочно-постоянной или кусочно-линейной аппроксимацией (см. рисунок):




В любом случае эффект неоптимальности может быть оценен посредством уравнения (25): кусочно-линейная форма (*г*) отличается от оптимальности в смысле уравнения (25) на несколько процентов (этот вывод может быть нарушен, скажем, из-за медленно убывающих «хвостов» распределения в случае бесконечных пределов интегрирования, так что может понадобиться вставка дополнительных линейных участков и т. д.).

Уравнение (27) позволяет говорить о неоптимальности моментов (т. е. их худшей информативности по сравнению с $f_{\text{opt}}$) на языке источников неоптимальности, т. е. тех отклонений $f_{\text{quasi}}(\mathbf{P})$ от $f_{\text{opt}}(\mathbf{P})$, которые дают доминирующие вклады в левую часть (24). Простейший пример: $f_{\text{opt}}$ — непрерывная плавно меняющаяся функция, тогда как

$f_{\text{quasi}}$ — кусочно-постоянная аппроксимация (см. рисунок, *в*); тогда $f_{\text{quasi}}$ сильнее всего отклоняется от $f_{\text{opt}}$ возле разрывов, которые поэтому естественно идентифицируются как источники неоптимальности. Тогда естественный способ улучшения $f_{\text{quasi}}$ — «регуляризация» разрывов посредством непрерывных (например, линейных) интерполяций.

Образно говоря, источники неоптимальности можно представлять себе как «дыры», через которые теряется информация о $M$, а регуляризация $f_{\text{quasi}}$ тогда будет соответствовать «затыканию» таких «дыр».

Если $\pi(\mathbf{P})$ дается теорией возмущений с возрастающе-сложными, но убывающе-важными вкладами, то может оказаться достаточным использовать приближенную форму для правой части (12), соответствующую учету лишь тех низших членов пертурбативного разложения, в которых проявляется зависимость от измеряемого параметра. Теоретические улучшения для полной плотности $\pi(\mathbf{P})$ не обязательно должны отражаться в квазиоптимальных моментах (разумеется, теоретическое среднее для квазиоптимального момента всегда должно вычисляться с помощью максимально полной теоретической плотности вероятности).

Если размерность пространства событий невелика, то может оказаться возможным построить подходящий $f_{\text{quasi}}$ методом грубой силы, генерируя гистограммы для $\pi(\mathbf{P})$ для разных $M$ и выполняя численное дифференцирование по $M$.

Можно также использовать для $f_{\text{quasi}}$ разные выражения: например, выполнить первые итерации с простым выражением для ускорения вычислений и в конце перейти на более сложную интерполяционную формулу для достижения наилучшей точности.

При оценивании нескольких параметров имеют место обычные неоднозначности, но всегда можно определить оптимальный момент для каждого параметра посредством (12). Тогда информативность (21) является матрицей (как и информация Фишера). Так как матрица корреляций (квази) оптимальных моментов известна (или может быть оценена из данных), то отображение эллипсоидов ошибок из пространства моментов в пространство параметров не составляет труда.

Популярный метод $\chi^2$ представляет собой фит с использованием ряда неоптимальных моментов (ячейки гистограммы). Гистограммирование приводит к потере информации, но зато метод является универсальным, проверяет форму распределения вероятностей и реализован в стандартных программах. С другой стороны, построение $f_{\text{quasi}}$ требует специальных усилий в каждой задаче, но зато потери информации могут быть в принципе сведены на нет подходящим выбором $f_{\text{quasi}}$. Баланс здесь, как обычно, — между качеством специальных решений и доступностью универсальных. Однако как только квазиоптимальные моменты построены, качество метода максимального правдоподобия оказывается доступным с меньшими вычислительными затратами. Два метода следует рассматривать как взаимно дополняющие: можно сначала использовать метод $\chi^2$ для верификации формы плотности вероятности и получения исходного значения $M_0$, которое затем используется для построения квазиоптимальных моментов с целью получения наилучшей окончательной оценки для $M$.

Дополнительное преимущество метода квазиоптимальных моментов может состоять в том, что некоторые из числа более изощренных теоретических формализмов могут давать предсказания для плотности вероятности в виде сингулярных (и поэтому не обязательно всюду положительно определенных) обобщенных функций (ср. систематическую калибровочно-инвариантную квантово-полевую теорию возмущений для нестабильных частиц, описанную в [7]). В таких случаях теоретические предсказания для обобщенных моментов могут существенно превосходить по качеству предсказания для плотности вероятности, так что использование метода $\chi^2$ может быть менее предпочтительным по сравнению с методом квазиоптимальных моментов.

Наконец, представляется реальным спроектировать более или менее универсальные программы для численного построения $f_{\mathrm{quasi}}$. В самом деле, задача построения $f_{\mathrm{quasi}}$ методом грубой силы с помощью генераторов событий близкородственна задаче об адаптивном интегрировании методом Монте-Карло, для которой имеется ряд алгоритмических схем [8]. Кроме того, наличие явных формул типа (12) и (20) позволяет использовать эвристики в тех сложных ситуациях (типа поиска топ-кварка в чисто адронном канале; см. [4]), в которых в настоящее время полагаются исключительно на метод проб и ошибок.

Автор благодарен Д. Ю. Бардину за обсуждение. Данная работа частично поддерживалась грантом РФФИ 99-02-18365.

## СПИСОК ЛИТЕРАТУРЫ

1. *Tkachov F. V.* A Theory of Jet Definition. hep-ph/9901444.

2. *Идье В. и др.* Статистические методы в экспериментальной физике: Пер. с англ. / Под ред. А. А. Тяпкина. М.: Атомиздат, 1976.

3. *Боровков А. А.* Математическая статистика. Оценивание параметров и проверка гипотез. М.: Наука, 1984.

4. *Bhat P. C., Prosper H., Snyder S. S.* Top Quark Physics at the Tevatron // Int. J. Mod. Phys. A. 1998. V. 13. P. 5113; hep-ex/9809011.

5. *Пытьев Ю. П., Шишмарев И. А.* Курс теории вероятностей и математической статистики для физиков. М.: Изд-во Моск. гос. ун-та, 1983.

6. Physics at LEP / Eds. J. Ellis, R. Peccei. Geneva: CERN, 1986.

7. *Ткачев Ф. В.* // Материалы V шк. ПИЯФ по теоретической физике / Под ред. Я. И. Азимова и др. Гатчина, 1999; hep-ph/9802307.

8. *Lepage G. P.* A New Algorithm for Adaptive Multidimensional Integration (VEGAS) // J. Comp. Phys. 1978. V. 27. P. 192.

9. *Manankova G. I., Tatarchenko A. F., Tkachov F. V.* MILXy Way: How Much Better than VEGAS Can One Integrate in Many Dimensions? Commun. FERMILAB-Conf-95/213-T.

10. *Jadach S.* Foam: Multi-Dimensional General Purpose Monte-Carlo Generator with Self-Adapting Simplectic Grid. physics/9910004.

Получено 5 апреля 2002 г.