

# Метод наименьших квадратов и линейная подгонка

---

## Содержание

---

|       |                                                   |     |
|-------|---------------------------------------------------|-----|
| 11.1. | Физическая задача: калибровка детектора . . . . . | 147 |
| 11.2. | От правдоподобия к сумме квадратов . . . . .      | 147 |
| 11.3. | Прямая по точкам . . . . .                        | 148 |
| 11.4. | Ковариация оценок . . . . .                       | 149 |
| 11.5. | Геометрия ошибок . . . . .                        | 150 |
| 11.6. | Возвращаемся к калибровке . . . . .               | 153 |
| 11.7. | Что остаётся после подгонки . . . . .             | 154 |
| 11.8. | Остатки и выбросы . . . . .                       | 156 |
| 11.9. | Задачи . . . . .                                  | 158 |
|       | Литература . . . . .                              | 159 |

**Аннотация**

В этой главе мы применим метод максимального правдоподобия к измерениям с гауссовыми ошибками и получим метод наименьших квадратов. На примере прямой найдём оценки параметров, их ошибки и корреляцию. Затем разберём, с какой точностью известна сама подогнанная кривая и что можно узнать из остатков. Физическим примером послужит калибровка энергетической шкалы детектора.

**11.1. Физическая задача: калибровка детектора**

Детектор выдаёт амплитуду сигнала в каналах аналого-цифрового преобразователя, сокращённо АЦП. Для физического анализа нужна энергия. Чтобы связать одно с другим, измерим положение нескольких пиков с известными энергиями.

Рассмотрим учебный набор из трёх измерений в таблице 11.1. Энергии считаем известными точно. Ошибки положений пиков независимы, гауссовы и имеют известное стандартное отклонение 2 канала. Здесь речь об ошибке измеренного центра пика; ширина самого пика может быть значительно больше.

**Таблица 11.1.** Учебные данные для калибровки энергетической шкалы

| Энергия $x_i$ , МэВ | Положение пика $y_i$ , канал | Ошибка $\sigma_i$ , канал |
|---------------------|------------------------------|---------------------------|
| 1                   | 103                          | 2                         |
| 2                   | 200                          | 2                         |
| 3                   | 303                          | 2                         |

Предположим, что в этом диапазоне средняя амплитуда линейно зависит от энергии:

$$f(x) = a + bx. \quad (11.1)$$

Параметр  $a$  задаёт смещение нуля в каналах, а  $b$  — коэффициент преобразования в каналах на МэВ. Подгонка, которую мы сейчас выведем, даёт

$$\hat{a} = (2.0 \pm 3.1) \text{ канала}, \quad \hat{b} = (100.0 \pm 1.4) \text{ канала/МэВ}. \quad (11.2)$$

Двух ошибок для описания результата ещё недостаточно. Оценки  $\hat{a}$  и  $\hat{b}$  сильно антикоррелированы: коэффициент корреляции равен примерно  $-0.93$ . Если увеличить наклон, свободный член придётся уменьшить, чтобы прямая по-прежнему проходила вблизи точек.

Откуда взялись эти числа? Почему коррелируют параметры, хотя ошибки трёх измерений независимы? И как эту корреляцию учитывать, когда калибровка уже используется в анализе? Начнём с функции правдоподобия.

**11.2. От правдоподобия к сумме квадратов**

Пусть имеются  $N$  измерений  $(x_i, y_i \pm \sigma_i)$  и модель  $y = f(x; \theta)$ . Во всей главе координаты  $x_i$  фиксированы и известны точно. Случайны результаты  $Y_i$ :

$$Y_i = f(x_i; \theta) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2). \quad (11.3)$$

Ошибки  $\varepsilon_i$  независимы, а все  $\sigma_i > 0$  известны и не зависят от параметров. Тогда правдоподобие имеет вид [1]

$$L(\theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{[y_i - f(x_i; \theta)]^2}{2\sigma_i^2}\right]. \quad (11.4)$$

Возьмём логарифм и умножим его на  $-2$ :

$$-2 \ln L(\theta) = \sum_{i=1}^N \frac{[y_i - f(x_i; \theta)]^2}{\sigma_i^2} + 2 \sum_{i=1}^N \ln(\sqrt{2\pi}\sigma_i). \quad (11.5)$$

Вторая сумма постоянна при изменении  $\theta$ . Максимум правдоподобия поэтому совпадает с минимумом функции

$$\chi^2(\theta) = \sum_{i=1}^N \frac{[y_i - f(x_i; \theta)]^2}{\sigma_i^2}. \quad (11.6)$$

Это и есть **взвешенный метод наименьших квадратов**, или МНК.

Разность между измеренным значением и моделью называется **остатком**. Обозначим её через  $r_i$ . Вес точки равен обратной дисперсии:

$$r_i(\theta) = y_i - f(x_i; \theta), \quad w_i = \frac{1}{\sigma_i^2}, \quad \chi^2 = \sum_i w_i r_i^2. \quad (11.7)$$

Каждый вклад в  $\chi^2$  безразмерен. Отклонение на два канала для точки с ошибкой один канал даёт вклад 4. Для точки с ошибкой четыре канала тот же остаток даёт всего 1/4. Так в подгонке учитывается разная точность измерений.

### 11.2.1. Что мы предположили об ошибках

Связь МНК с максимумом правдоподобия получена для конкретной модели 11.3. Если  $\sigma_i$  зависит от параметров, вторая сумма в формуле 11.5 тоже меняется, и отбрасывать её нельзя. При малом числе событий в бине следует использовать пуассоновское правдоподобие. Подстановка  $\sigma_i = \sqrt{y_i}$  особенно неудачна для пустого бина: его вес тогда вообще не определён.

Сам метод наименьших квадратов можно применять и к негауссовым ошибкам. Ниже мы увидим, что несмещённость линейных оценок и формулы их ковариации следуют из первых двух моментов ошибок. Но совпадение с методом максимального правдоподобия и точные гауссовы интервалы в общем случае теряются.

Неопределённости по  $x$  и корреляции между измерениями требуют другого описания данных. В этой главе их пока нет. Физические ограничения на параметры тоже нужно учитывать отдельно: минимум может оказаться на границе, где обычные симметричные ошибки уже не описывают интервал.

## 11.3. Прямая по точкам

Вернёмся к модели  $f(x) = a + bx$ . Нужно минимизировать

$$\chi^2(a, b) = \sum_i w_i (y_i - a - bx_i)^2. \quad (11.8)$$

Слово «линейная» относится к зависимости от неизвестных параметров. Например,  $a + bx + cx^2$  тоже линейная модель по  $a, b, c$ , хотя её график — парабола. Для начала достаточно двух параметров прямой.

### 11.3.1. Нормальные уравнения

Приравняем производные к нулю:

$$\begin{aligned}\frac{\partial \chi^2}{\partial a} &= -2 \sum_i w_i (y_i - a - bx_i) = 0, \\ \frac{\partial \chi^2}{\partial b} &= -2 \sum_i w_i x_i (y_i - a - bx_i) = 0.\end{aligned}\quad (11.9)$$

Введём пять сумм. Все они вычисляются по данным и известным весам:

$$\begin{aligned}S &= \sum_i w_i, & S_x &= \sum_i w_i x_i, & S_y &= \sum_i w_i y_i, \\ S_{xx} &= \sum_i w_i x_i^2, & S_{xy} &= \sum_i w_i x_i y_i.\end{aligned}\quad (11.10)$$

Тогда два условия 11.9 образуют систему

$$\begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}.\quad (11.11)$$

Это **нормальные уравнения** МНК. Шляпки у параметров появились потому, что мы уже ищем их оценки. Обозначим матрицу слева через  $A$ , вектор параметров через  $\beta = (a, b)^T$ , а правую часть через  $\mathbf{c}$ . Решение записывается как  $\hat{\beta} = A^{-1}\mathbf{c}$ .

Для матрицы размера  $2 \times 2$  обратную матрицу можно выписать сразу. С определителем  $\Delta = SS_{xx} - S_x^2$  получаем

$$\hat{a} = \frac{S_{xx}S_y - S_xS_{xy}}{\Delta}, \quad \hat{b} = \frac{SS_{xy} - S_xS_y}{\Delta}.\quad (11.12)$$

Численный поиск минимума здесь не требуется. При большом числе линейных параметров систему обычно решают средствами линейной алгебры, без явного вычисления обратной матрицы.

### 11.3.2. Можно ли определить наклон?

Введём взвешенный центр координат и меру их разброса:

$$\bar{x}_w = \frac{S_x}{S}, \quad Q_x = \sum_i w_i (x_i - \bar{x}_w)^2 = S_{xx} - \frac{S_x^2}{S} = \frac{\Delta}{S}.\quad (11.13)$$

При положительных весах  $Q_x$  обращается в нуль только тогда, когда все  $x_i$  одинаковы. В этом случае мы измеряем значение прямой в одной точке. Через неё можно провести прямые с любым наклоном, меняя свободный член. Данные позволяют определить одну комбинацию  $a + bx_1$ , а двух отдельных параметров не определяют.

Если есть хотя бы два различных  $x_i$ , то  $Q_x > 0$  и  $\Delta > 0$ . Матрица  $A$  положительно определена, поэтому найденная стационарная точка является единственным минимумом.

## 11.4. Ковариация оценок

В прошлой главе ошибка определялась по кривизне правдоподобия. Для прямой с известными гауссовыми ошибками этот способ даёт точный результат при любом числе точек, позволяющем определить оба параметра.

Положим  $\delta a = a - \hat{a}$  и  $\delta b = b - \hat{b}$ . Раскрывая квадраты в формуле 11.8, получаем

$$\chi^2(a, b) - \chi_{\min}^2 = S(\delta a)^2 + 2S_x \delta a \delta b + S_{xx}(\delta b)^2. \quad (11.14)$$

Линейные члены исчезли благодаря нормальным уравнениям. Членов третьей и более высокой степени нет: исходная функция уже была квадратичной. Следовательно, относительное правдоподобие равно

$$\frac{L(\beta)}{L(\hat{\beta})} = \exp\left[-\frac{1}{2}(\beta - \hat{\beta})^T A(\beta - \hat{\beta})\right]. \quad (11.15)$$

Матрица  $A$  играет роль обратной ковариации. Проверим это непосредственно по распределению оценок, чтобы не приписывать самому правдоподобию смысл распределения параметров.

#### Вывод

##### 11.4.1. Дисперсии и ковариация из случайных ошибок

Вектор  $\mathbf{c}$  из нормальных уравнений содержит две случайные суммы:  $S_y = \sum_i w_i Y_i$  и  $S_{xy} = \sum_i w_i x_i Y_i$ . Подстановка  $Y_i = a + bx_i + \epsilon_i$  даёт

$$\begin{aligned} \mathbb{E}[\mathbf{c}] &= A\beta, \\ \text{Cov}(\mathbf{c}) &= \begin{pmatrix} \sum_i w_i^2 \sigma_i^2 & \sum_i w_i^2 x_i \sigma_i^2 \\ \sum_i w_i^2 x_i \sigma_i^2 & \sum_i w_i^2 x_i^2 \sigma_i^2 \end{pmatrix} = A. \end{aligned} \quad (11.16)$$

Использовано равенство  $w_i \sigma_i^2 = 1$  и отсутствие ковариаций между ошибками разных точек. Поскольку  $\hat{\beta} = A^{-1}\mathbf{c}$ ,

$$\mathbb{E}[\hat{\beta}] = \beta, \quad V_{\beta} = \text{Cov}(\hat{\beta}) = A^{-1}A(A^{-1})^T = A^{-1}. \quad (11.17)$$

Обращение матрицы даёт

$$V_{\beta} = \frac{1}{\Delta} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}. \quad (11.18)$$

Итак, оценки несмещённые. Их дисперсии и ковариация — три элемента матрицы 11.18:

$$\text{Var}(\hat{a}) = \frac{S_{xx}}{\Delta}, \quad \text{Var}(\hat{b}) = \frac{S}{\Delta}, \quad \text{cov}(\hat{a}, \hat{b}) = -\frac{S_x}{\Delta}. \quad (11.19)$$

В этом выводе понадобились только средние и ковариации ошибок. Гауссова модель дополнительно гарантирует гауссово распределение вектора оценок: он является линейной функцией гауссовых данных.

Обратите внимание: в матрице  $V_{\beta}$  вообще нет значений  $y_i$ . При фиксированных  $x_i$  и известных  $\sigma_i$  случайные значения  $y_i$  меняют положение минимума, но его кривизна остаётся той же. Это свойство именно линейной модели с заданными ошибками.

## 11.5. Геометрия ошибок

Через  $\bar{x}_w$  и  $Q_x$  формулы 11.19 переписываются в виде

$$\sigma_b^2 = \frac{1}{Q_x}, \quad \sigma_a^2 = \frac{1}{S} + \frac{\bar{x}_w^2}{Q_x}, \quad \text{cov}(\hat{a}, \hat{b}) = -\frac{\bar{x}_w}{Q_x}. \quad (11.20)$$

Здесь  $\sigma_a$  и  $\sigma_b$  обозначают стандартные отклонения оценок. Чем шире разнесены точки по  $x$  при тех же ошибках, тем больше  $Q_x$  и тем точнее определяется наклон. Для

калибровки это означает: пики в узком энергетическом диапазоне плохо определяют коэффициент преобразования.

Со свободным членом ситуация другая. Параметр  $a$  — значение прямой при  $x = 0$ . Если калибровочные точки находятся далеко от нуля, приходится экстраполировать к нему прямую. Неопределённость наклона увеличивает неопределённость такого пересечения.

### 11.5.1. Через какую точку проходит прямая

Зафиксируем произвольный наклон  $b$  и найдём лучшее значение  $a$ . Первое нормальное уравнение даёт

$$\hat{a}(b) = \bar{y}_w - b\bar{x}_w, \quad \bar{y}_w = \frac{S_y}{S}. \quad (11.21)$$

Двойная шляпка имеет тот же смысл, что и в прошлой главе: свободный член подобран при фиксированном наклоне. Подставляя  $x = \bar{x}_w$ , получаем  $f(\bar{x}_w) = \bar{y}_w$ . Все такие прямые проходят через взвешенный центр данных  $(\bar{x}_w, \bar{y}_w)$ .

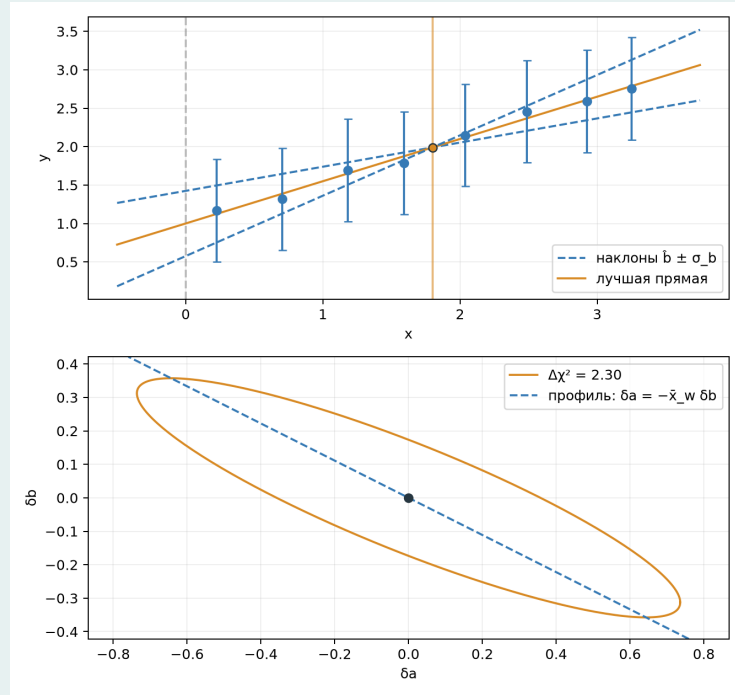
Теперь понятен знак ковариации. Если  $\bar{x}_w > 0$ , увеличение наклона компенсируется уменьшением свободного члена. При  $\bar{x}_w < 0$  оба параметра меняются в одну сторону. Коэффициент корреляции равен

$$\rho_{ab} = \frac{\text{cov}(\hat{a}, \hat{b})}{\sigma_a \sigma_b} = -\frac{S_x}{\sqrt{SS_{xx}}}. \quad (11.22)$$

### 11.5.2. Интерактив: центр точек и наклон

В апплете можно менять взвешенный центр  $\bar{x}_w$  и среднеквадратичный разброс координат  $s_x = \sqrt{Q_x/S}$ . Суммарный вес  $S = 18$  остаётся фиксированным. Координаты здесь безразмерные. Учебные точки построены так, чтобы их лучшая прямая была  $y = 1 + 0.55x$ .

Сплошная оранжевая линия показывает эту прямую. Две синие пунктирные линии соответствуют наклонам  $\hat{b} \pm \sigma_b$  с заново подобранным свободным членом. На втором графике изображён контур  $\Delta\chi^2 = 2.30$  и направление профилирования  $\delta a = -\bar{x}_w \delta b$ .



### Интерактив. Геометрия корреляции параметров

При  $\bar{x}_w = 1.8$  и  $s_x = 1$  корреляция оценок равна  $-0.87$ . Перенос центра к нулю устраняет наклон эллипса. Увеличение разброса точек уменьшает ошибку наклона прямой.



#### 11.5.3. Центрированная параметризация

Можно сразу измерять  $x$  от центра данных:

$$x' = x - \bar{x}_w, \quad f(x) = a' + bx', \quad a' = a + b\bar{x}_w. \quad (11.23)$$

Теперь  $\sum_i w_i x'_i = 0$ , поэтому

$$\hat{a}' = \bar{y}_w, \quad \text{Var}(\hat{a}') = \frac{1}{S}, \quad \text{cov}(\hat{a}', \hat{b}) = 0. \quad (11.24)$$

Наклон и его ошибка сохранились. Параметр  $a'$  теперь описывает значение прямой в центре данных. Сама калибровка не изменилась, но её параметры стали некоррелированными. В гауссовой модели оценки  $\hat{a}'$  и  $\hat{b}$  также независимы.

#### 11.5.4. Нормальные координаты и профиль

В новых координатах квадратичная форма диагональна:

$$\Delta\chi^2 = S(\delta a')^2 + Q_x(\delta b)^2 = z_a^2 + z_b^2, \quad z_a = \sqrt{S} \delta a', \quad z_b = \sqrt{Q_x} \delta b. \quad (11.25)$$

Мы получили те же нормальные координаты, которые вводили для двумерного гаусса. Эллипсы превратились в окружности. Здесь преобразование точно во всей плоскости параметров.

При профилировании по  $a'$  первое слагаемое обращается в нуль. Остаётся

$$\chi_p^2(b) - \chi_{\min}^2 = Q_x(b - \hat{b})^2 = \frac{(b - \hat{b})^2}{\sigma_b^2}. \quad (11.26)$$

Поэтому пределы  $\hat{b} \pm \sigma_b$  соответствуют росту профильного  $\chi^2$  на единицу. Это одномерный интервал с покрытием около 68.3% в нашей гауссовой модели. Для совместной области двух параметров с тем же покрытием нужен порог  $\Delta\chi^2 = 2.30$ . Именно он показан в апплете.

## 11.6. Возвращаемся к калибровке

Для данных из таблицы 11.1 все веса равны  $1/4$  в обратных квадратах канала. Получаем

$$\bar{x}_w = 2 \text{ МэВ}, \quad \bar{y}_w = 202 \text{ канала}, \quad Q_x = 0.5 \text{ МэВ}^2/\text{канал}^2. \quad (11.27)$$

Из формул 11.12 и 11.20 следуют  $\hat{b} = 100$  каналов/МэВ и  $\hat{a} = 2$  канала,  $\sigma_b = \sqrt{2}$  канала/МэВ и  $\sigma_a = \sqrt{28/3}$  канала. Ковариация равна  $-4$  канал<sup>2</sup>/МэВ. Это и даёт результат 11.2 с корреляцией  $-\sqrt{6/7} \simeq -0.93$ .

В центрированной записи та же прямая выглядит проще:

$$\hat{f}(x) = 202 \text{ канала} + \left(100 \frac{\text{каналов}}{\text{МэВ}}\right)(x - 2 \text{ МэВ}). \quad (11.28)$$

Значение в центре равно  $202 \pm 1.15$  канала и не коррелирует с наклоном. Ошибка 1.15 заметно меньше ошибки свободного члена 3.1: центр данных находится внутри диапазона измерений, а нулевая энергия — за его пределами.

### 11.6.1. С какой точностью известна прямая?

В произвольной фиксированной точке  $x$  оценка среднего сигнала равна  $\hat{f}(x) = \hat{a} + \hat{b}x$ . Закон распространения ошибок даёт

$$\begin{aligned} \text{Var}[\hat{f}(x)] &= \sigma_a^2 + 2x \text{cov}(\hat{a}, \hat{b}) + x^2 \sigma_b^2 \\ &= \frac{1}{S} + \frac{(x - \bar{x}_w)^2}{Q_x}. \end{aligned} \quad (11.29)$$

Это точное равенство:  $\hat{f}$  линейна по оценкам. Минимальная ошибка достигается в центре данных. По мере удаления от центра полоса вокруг подогнанной прямой расширяется.

Если забыть ковариацию, уже при  $x = 2$  МэВ получится дисперсия  $28/3 + 8 = 52/3$  канал<sup>2</sup> вместо  $4/3$  канал<sup>2</sup>. Ошибка вырастет с 1.15 до 4.16 канала. Сильная корреляция здесь имеет вполне измеримое последствие.

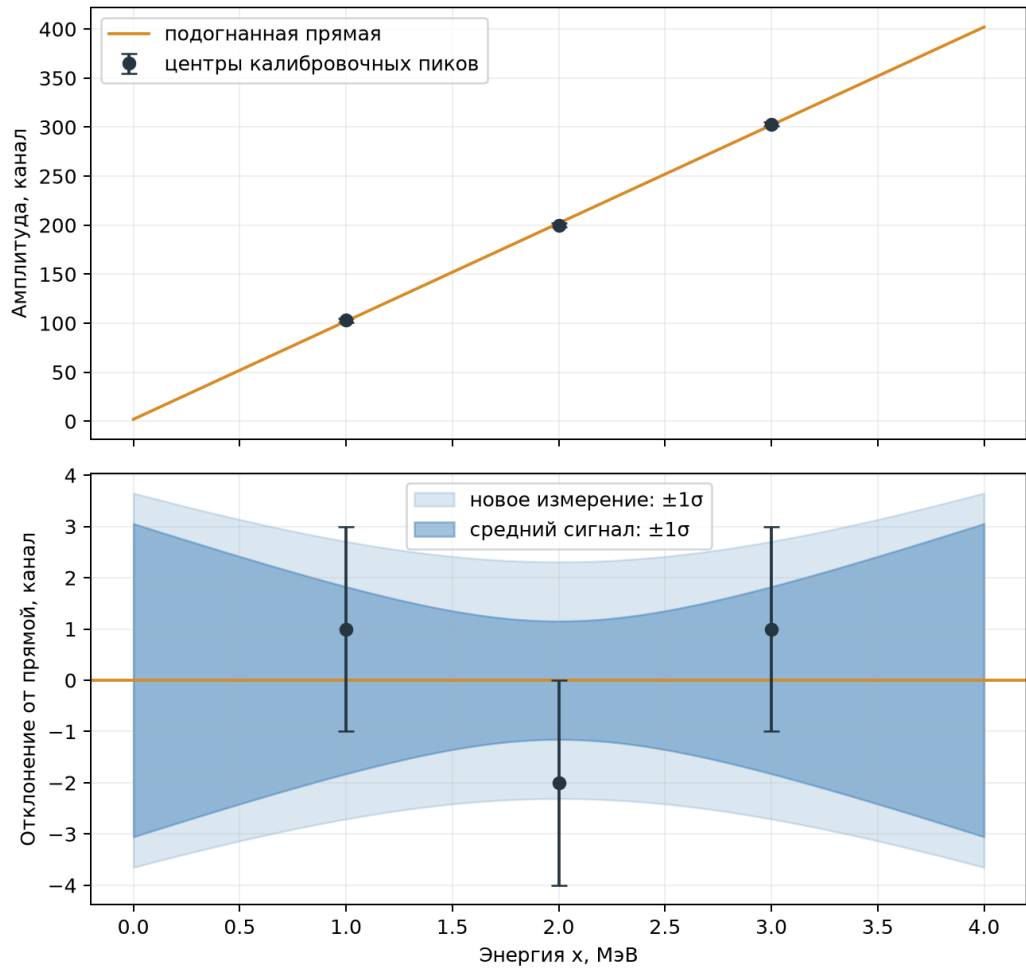
### 11.6.2. Средний сигнал и новое измерение

Теперь проведём ещё одно независимое измерение при фиксированной энергии  $x_*$ , с гауссовой ошибкой известной ширины  $\sigma_*$ :

$$Y_* = a + bx_* + \varepsilon_*, \quad \text{Var}[Y_* - \hat{f}(x_*)] = \sigma_*^2 + \frac{1}{S} + \frac{(x_* - \bar{x}_w)^2}{Q_x}. \quad (11.30)$$

Первое слагаемое описывает флуктуацию нового измерения. Два остальных связаны с неопределённостью калибровочной прямой. Поэтому интервал для нового измерения шире интервала для среднего сигнала.

В центре нашей калибровки средний сигнал известен с ошибкой 1.15 канала. Если новое измерение имеет  $\sigma_* = 2$  канала, стандартное отклонение разности  $Y_* - \hat{f}(x_*)$  равно  $\sqrt{4 + 4/3} = 2.31$  канала. Обе полосы показаны на рисунке 11.1.



**Рис. 11.1.** Учебная калибровка из таблицы: точки с ошибками и подогнанная прямая. На нижнем графике из каждой величины вычтена подогнанная прямая, чтобы различить полосы. Узкая полоса соответствует одной стандартной ошибке среднего сигнала; широкая учитывает также ошибку нового независимого измерения  $\sigma_* = 2$  канала.

Границы обеих полос имеют точечное покрытие около 68.3% при каждом заранее фиксированном  $x$ . Утверждение о всей кривой сразу требует одновременной доверительной полосы с другим порогом. Кроме того, за пределами калибровочного диапазона эти формулы учитывают лишь ошибки параметров прямой. Возможную нелинейность реального отклика они не описывают.

### 11.7. Что остаётся после подгонки

Остатки при найденных параметрах равны

$$r_i = y_i - \hat{a} - \hat{b}x_i, \quad u_i = \frac{r_i}{\sigma_i}. \quad (11.31)$$

Величины  $u_i$  позволяют сравнивать отклонения точек с разными ошибками. Однако считать их независимой выборкой из  $\mathcal{N}(0, 1)$  нельзя. Прямая уже подстроилась под те же точки. Из нормальных уравнений следуют два ограничения:

$$\sum_i w_i r_i = 0, \quad \sum_i w_i x_i r_i = 0. \quad (11.32)$$

Изменение одной точки меняет прямую и тем самым остатки остальных точек.

Найдём эту связь явно.

### 11.7.1. Матрица влияния

Разделим данные и модель в каждой строке на  $\sigma_i$ :

$$z_i = \frac{y_i}{\sigma_i}, \quad B_{i1} = \frac{1}{\sigma_i}, \quad B_{i2} = \frac{x_i}{\sigma_i}. \quad (11.33)$$

Матрица  $B$  имеет  $N$  строк и два столбца. Ошибки вектора  $\mathbf{z}$  независимы и имеют единичную дисперсию. Нормальные уравнения теперь записываются как  $B^T B \hat{\beta} = B^T \mathbf{z}$ . Значения подогнанной модели в этих координатах равны

$$\hat{\mathbf{z}} = H\mathbf{z}, \quad H = B(B^T B)^{-1} B^T. \quad (11.34)$$

Матрица  $H$  называется **матрицей влияния**. Она симметрична и удовлетворяет  $H^2 = H$ : если ещё раз подогнать уже лежащие на прямой значения, они не изменятся. Вектор нормированных остатков равен  $\mathbf{u} = (I - H)\mathbf{z}$ , где  $I$  — единичная матрица. Поэтому

$$\text{Cov}(\mathbf{u}) = (I - H)(I - H)^T = I - H. \quad (11.35)$$

В частности,

$$\text{Var}(u_i) = 1 - h_{ii}, \quad h_{ii} = w_i \left[ \frac{1}{S} + \frac{(x_i - \bar{x}_w)^2}{Q_x} \right]. \quad (11.36)$$

Точка с большим  $h_{ii}$  сильнее подтягивает к себе прямую, поэтому её остаток имеет меньшую дисперсию. При прочих равных большое влияние имеют точки далеко от центра  $x$ . Малый остаток такой точки сам по себе ещё не означает, что с измерением всё в порядке.

Для известной дисперсии ошибки и  $h_{ii} < 1$  величина  $u_i / \sqrt{1 - h_{ii}}$  имеет единичную дисперсию, а в нашей гауссовой модели и стандартное нормальное распределение. Эта нормировка не устраняет корреляций между разными остатками.

### 11.7.2. Почему остаётся $N - 2$ степеней свободы

Матрица  $H$  проектирует данные на двумерное пространство моделей, заданное двумя столбцами  $B$ . У неё два собственных значения 1, остальные равны 0, и  $\text{tr } H = 2$ . Для остатков остаётся  $N - 2$  независимых направлений. При верной линейной модели и известных гауссовых ошибках

$$\chi_{\min}^2 = \mathbf{u}^T \mathbf{u} \sim \chi_{N-2}^2, \quad \mathbb{E}[\chi_{\min}^2] = N - 2. \quad (11.37)$$

Распределение справа обозначено тем же символом  $\chi^2$ , что и функция, которую мы минимизировали. Здесь  $N > 2$ . Для нашей калибровки остатки равны 1, -2, 1 канала, поэтому  $\chi_{\min}^2 = 1.5$  при одной степени свободы. Из одного такого числа нельзя заключить, что модель детектора верна. Формальную проверку согласия разберём в главе 13.

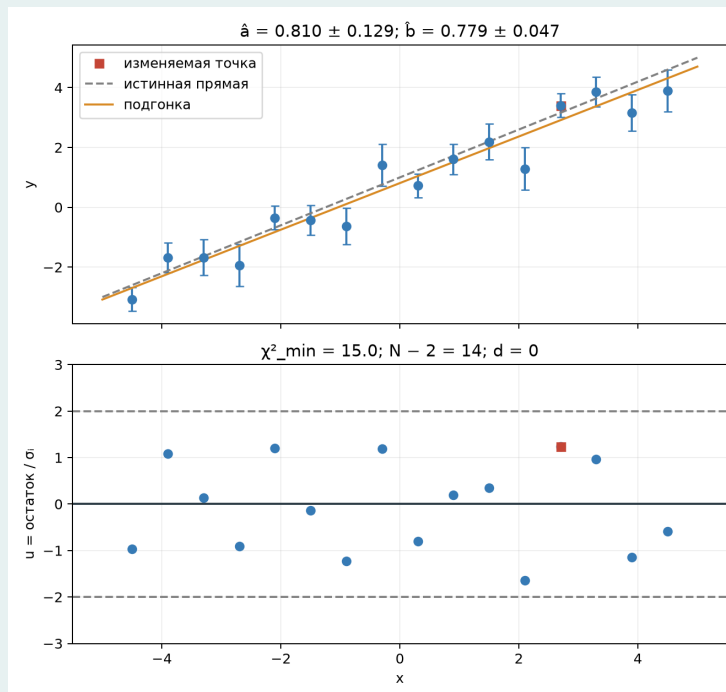
Если ошибки всех точек имеют одинаковую, но неизвестную дисперсию  $\sigma^2$ , её можно оценить по остаткам:

$$s^2 = \frac{1}{N - 2} \sum_i r_i^2, \quad \mathbb{E}[s^2] = \sigma^2. \quad (11.38)$$

Для интервалов тогда требуется учитывать случайность  $s$ , что приводит к распределению Стьюдента. В нашей основной задаче все  $\sigma_i$  заданы заранее. Перемасштабировать их до  $\chi_{\min}^2 / (N - 2) = 1$  оснований нет.

## 11.8. Остатки и выбросы

Сгенерируем 16 точек около прямой  $y = 1 + 0.8x$  с независимыми гауссовыми ошибками. Обе координаты здесь безразмерные. В апплете можно изменить масштаб ошибок и добавить сдвиг  $d$  к одной отмеченной точке. При изменении сдвига остальные точки сохраняются. Кнопка «Новые данные» меняет всю случайную выборку.

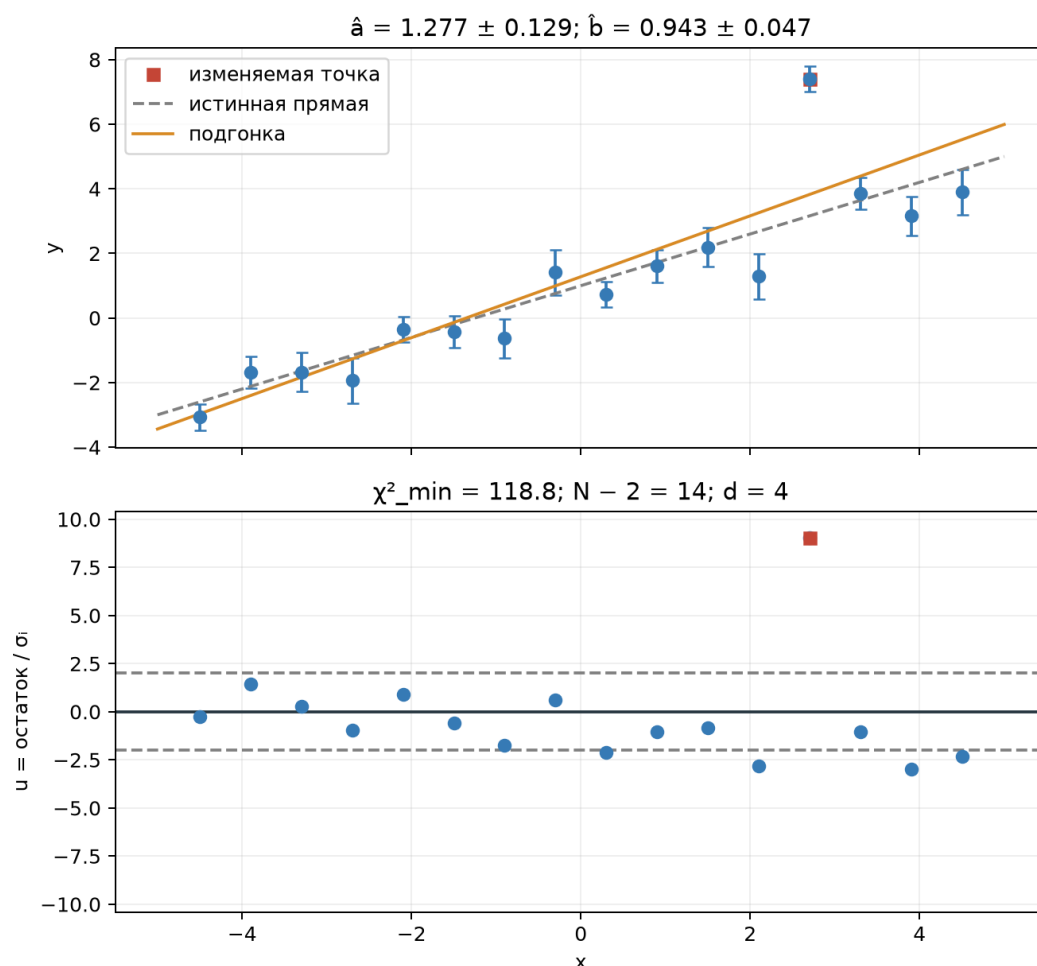


### Интерактив. Прямая, остатки и выброс

Подгонка 16 псевдоизмерений при масштабе ошибок 0.55 и без дополнительного сдвига. Пунктиром показана истинная прямая, сплошной линией — результат подгонки. Нижний график содержит остатки, делённые на ошибки точек.



На рисунке 11.2 к той же выборке добавлен сдвиг  $d = 4$  в одной точке. Её заявленная ошибка не изменилась. Подгонка учитывает эту точку с прежним весом, и прямая смещается.



**Рис. 11.2.** Те же псевдоизмерения после сдвига отмеченной точки на  $d = 4$ . Пунктирная прямая задаёт исходную модель, сплошная получена из подгонки. Сдвиг одной точки изменил и её собственный остаток, и остатки остальных измерений.

Этот эффект можно вычислить без повторной минимизации. Если заменить  $y_j$  на  $y_j + d$  при неизменных координатах и весах, то

$$\delta \hat{b} = \frac{w_j(x_j - \bar{x}_w)}{Q_x} d, \quad \delta \hat{a} = \frac{w_j}{S} d - \bar{x}_w \delta \hat{b}. \quad (11.39)$$

Смещение наклона растёт линейно с  $d$  и с расстоянием точки от центра. Обычный МНК не ограничивает влияние сколь угодно большого отклонения.

Что делать, если такая точка встретила в эксперименте? Нужно проверить причину: сбой электроники, неверную ошибку измерения, примесь другого процесса, насыщение детектора или недостаточность самой модели. Удаление точки требует физического основания и явно описанного правила отбора. Если большие отклонения составляют часть распределения ошибок, это распределение следует включить в правдоподобие.

На графике остатков полезно смотреть также на согласованное поведение многих точек. Изгиб может указывать на нелинейность отклика. Рост разброса с  $x$  может означать, что зависимость ошибок от энергии описана неверно. Такой график помогает поставить вопрос к модели; сам по себе он ещё не выбирает причину отклонения.

**Итоги главы**

- Для независимых гауссовых измерений с известными, не зависящими от параметров дисперсиями МНК совпадает с методом максимального правдоподобия.
- Подгонка прямой сводится к двум нормальным уравнениям. Ковариация оценок определяется координатами точек и их ошибками.
- Центрирование координаты устраняет ковариацию наклона и значения прямой в центре данных. При вычислении ошибки кривой в исходных параметрах ковариационный член необходимо сохранять.
- После подгонки остатки связаны между собой и имеют уменьшенные дисперсии. Их структура и влияние отдельных точек позволяют проверять описание измерений.

**11.9. Задачи****Задача 1. Среднее как результат МНК**

Имеются независимые измерения одной величины:

$$Y_i \sim N(\mu, \sigma_i^2).$$

Все  $\sigma_i > 0$  известны и не зависят от  $\mu$ .

1. Запишите функцию  $\chi^2(\mu)$ .
2. Найдите её минимум.
3. Покажите, что полученная оценка является взвешенным средним.
4. Найдите дисперсию этой оценки.
5. Рассмотрите частный случай одинаковых  $\sigma_i$ .

**Задача 2. Аналитическая подгонка прямой**

Для данных  $(x_i, y_i, \sigma_i)$  рассматривается модель

$$y = a + bx.$$

Координаты  $x_i$  известны точно, ошибки по  $y$  независимы и гауссовы, их стандартные отклонения  $\sigma_i > 0$  известны. Среди  $x_i$  есть хотя бы два различных значения.

1. Выведите нормальные уравнения.
2. Получите аналитические формулы для  $\hat{a}$  и  $\hat{b}$ .
3. Выведите ковариационную матрицу параметров.
4. Покажите, что центрирование  $x_i$  около взвешенного среднего устраняет ковариацию параметров.
5. Проверьте формулы численно на искусственной выборке.

**Задача 3. Ошибка подогнанной кривой**

Три независимых гауссовых измерения при точно известных энергиях  $x = (1, 2, 3)$  МэВ дали положения пиков  $y = (103, 200, 303)$  канала. Ошибка каждого положения равна 2 каналам. Для модели  $f(x) = a + bx$  найдите:

1. дисперсию оценки среднего  $\hat{y}(x) = \hat{a} + \hat{b}x$ ;
2. точку, в которой эта дисперсия минимальна;
3. дисперсию разности  $Y_* - \hat{y}(x_*)$  для нового независимого измерения при фиксированном  $x_*$  с известной ошибкой  $\sigma_*$ ;
4. различие между полосой среднего сигнала и полосой для нового измерения.

Постройте обе полосы для этих данных при  $\sigma_* = 2$  канала. Объясните, почему точечное покрытие не означает покрытия всей кривой сразу.

**Задача 4. Остатки и диагональные элементы матрицы влияния**

Пусть  $N > 2$ ,  $y_i = a + bx_i + \varepsilon_i$ , а  $\varepsilon_i$  независимы, имеют нулевые средние и одну известную дисперсию  $\sigma^2$ . Координаты  $x_i$  фиксированы и не все одинаковы. В этом случае  $\hat{\mathbf{y}} = H\mathbf{y}$ , где  $H = X(X^T X)^{-1} X^T$ ,  $X_{i1} = 1$ ,  $X_{i2} = x_i$ . Обозначьте остатки через  $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ , а единичную матрицу через  $I$ .

1. покажите, что  $H^2 = H$ ;
2. покажите, что  $\text{tr } H = 2$ ;
3. выведите ковариацию остатков

$$\text{Cov}(\mathbf{r}) = \sigma^2(I - H);$$

4. найдите диагональные элементы  $h_{ii}$ , характеризующие влияние точек при подгонке прямой;
5. добавьте одну далёкую по  $x$  точку и исследуйте её влияние.

**Задача 5. Влияние одной точки**

Сгенерируйте 16 независимых гауссовых измерений около прямой  $y = 1 + 0.8x$  при  $x_i = -4.5 + 9(i - 1)/15$ ,  $i = 1, \dots, 16$ , и известных одинаковых ошибках  $\sigma_i = 0.55$ . Зафиксируйте начальное состояние генератора.

1. Выполните подгонку обычным методом наименьших квадратов.
2. Добавьте к одной точке сдвиг  $d$ , сохранив её заявленную ошибку. Выведите изменения  $\hat{a}$  и  $\hat{b}$  из нормальных уравнений.
3. Меняйте  $d$  от  $-4$  до  $6$  и сравните формулы с повторной подгонкой.
4. Повторите опыт для точки вблизи центра и для крайней точки. Сравните влияние на наклон и остатки.
5. Проверьте, меняется ли вычисленная ковариация параметров при этих сдвигах. Объясните, почему неизменная ошибка параметра не гарантирует отсутствия смещения из-за испорченного измерения.

**Литература**

[1] Cowan, G. *Statistical Data Analysis*. Oxford University Press. 1998.