Осень 2021 г.

Специальный практикум (V курс)

Разложение бимодальной функции распределения на две нормальные составляющие методом наибольшего правдоподобия

И. И. Никифоров

1. Метод наибольшего правдоподобия и метод наименьших квадратов

МНК — стандартный метод оценивания неизвестных параметров в случаях, когда наблюдаемые величины (измерения) можно сопоставить с модельными предсказаниями, например

$$V_{\rm obs} = V_{\rm mod}(l, b, r; \mathbf{a}). \qquad (1)$$

Тогда, предполагая, что случайная величина $V_{\rm obs} - V_{\rm mod}$ распределена по нормальному закону $\mathcal{N}(0, \sigma^2)$, можно найти параметры, минимизируя сумму квадратов невязок. Метод хорошо всем известен, особенно в линейном случае.

Но так бывает не всегда. Бывает, что распределение случайных (измеряемых) величин заведомо нельзя описать как одномерное нормальное. Не всегда вообще удается написать систему уравнений (1). Примеры: нахождение дисперсионных характеристик (параметров эллипсоида скоростей), изучение сразу двух и более галактических подсистем. В этих случаях, если вид распределения (ненормального) известен, можно применить *метод наибольшего правдоподобия* (maximum likelihood estimation, MLE).

Простейший пример — оценка параметров $a_1, a_2, \ldots, a_M$ распределения одномерного аргумента x с плотностью вероятности

$$\varphi(x; a_1, a_2, \ldots, a_M). \qquad (2)$$

Пусть получена случайная выборка значений аргумента $x_1, x_2, \ldots, x_N$. Плотность вероятности для этой выборки равна

$$L(x_1, x_2, \ldots, x_N; a_1, a_2, \ldots, a_M) = \prod_{i=1}^{N} \varphi(x_i; a_1, a_2, \ldots, a_M). \qquad (3)$$

Функция L называется *функцией правдоподобия*. По сути, это — вероятность получить именно такую комбинацию измерений (наблюдений) $x_1, x_2, \ldots, x_N$ (вероятность попадания в соответствующий дифференциально малый гиперкуб) для одномерной плотности вероятности (2).

Принцип наибольшего правдоподобия состоит в том, что выбираются такие значения $a_{1,0}, a_{2,0},\ \ldots, a_{M,0}$, при которых функция L достигает максимума.

$$\mathbf{a}_0: \quad L(\mathbf{x};\mathbf{a}_0) = \max\,. \tag{4}$$

Эти значения называют *точечными оценками* параметров $a_1, a_2,\ \ldots, a_M$.

Для практического решения задачи вместо L удобно рассматривать логарифмическую функцию правдоподобия

$$\mathcal{L} \equiv -\ln L = -\sum_{i=1}^{N} \ln\varphi(x_i; a_1, a_2,\ \ldots, a_M). \tag{5}$$

Для заданной так логарифмической функции правдоподобия точечные оценки параметров дает ее *минимум*:

$$\mathbf{a}_0: \quad \mathcal{L}(\mathbf{x};\mathbf{a}_0) = \min\,. \tag{6}$$

Способы поиска точки минимума $\mathcal{L}$.

1. Решение системы уравнений

$$\frac{\partial\ln\mathcal{L}}{\partial a_j} = -\sum_{i=1}^{N} \frac{\partial\ln\varphi(x_i; a_1, a_2,\ \ldots, a_M)}{\partial a_j} = 0, \qquad j = 1, 2,\ \ldots, M. \tag{7}$$

Пример 1. Нормальное распределение:

$$\varphi(x; a_1, a_2) = \mathcal{N}(\bar{x}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\dfrac{(x-\bar{x})^2}{2\sigma^2}}\,. \tag{8}$$

Решение системы (7) в этом случае дает

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2. \tag{9}$$

Пример 2. $V_{\rm obs} - V_{\rm mod}(\mathbf{a}) \in \mathcal{N}(0, \sigma^2)$. Решение системы (7) в этом случае дает МНК:

$$\sum_{i=1}^{N} [V_{\rm obs} - V_{\rm mod}(\mathbf{a})]_i^2 \to \min\,. \tag{10}$$

2. Численный поиск минимума.

2. Разложение бимодального распределения на две нормальные составляющие

Рассмотрим распределение металличностей шаровых скоплений Галактики, о котором установлено, что оно является бимодальным. Из наблюдений известны измерения металличностей отдельных скоплений

$$f_i \equiv [\mathrm{Fe/H}]_i = \left[\lg\left(\frac{\mathrm{Fe}}{\mathrm{H}}\right)_* - \lg\left(\frac{\mathrm{Fe}}{\mathrm{H}}\right)_\odot\right]_i, \quad i = 1, \ldots, N. \tag{11}$$

Задача: разложить наблюдаемое распределение на две нормальные составляющие

$$\mathcal{N}_1(\bar{F}_1, \sigma_1^2), \quad \mathcal{N}_2(\bar{F}_2, \sigma_2^2). \tag{12}$$

Тогда имеем следующий модельный дифференциальный закон распределения:

$$\varphi(f) = \frac{c}{\sqrt{2\pi}\sigma_1} e^{-\dfrac{(f-\bar{F}_1)^2}{2\sigma_1^2}} + \frac{1-c}{\sqrt{2\pi}\sigma_2} e^{-\dfrac{(f-\bar{F}_2)^2}{2\sigma_2^2}}. \tag{13}$$

Функция правдоподобия:

$$L(f_1, f_2, \ldots, f_N; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c) = \prod_{i=1}^{N} \varphi(f_i; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c). \tag{14}$$

Логарифмическая функция правдоподобия:

$$\begin{aligned}\mathcal{L} &= -\sum_{i=1}^{N} \ln \varphi(f_i; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c) = \\ &= \frac{N}{2}\ln(2\pi) - \sum_{i=1}^{N} \ln\left[\frac{c}{\sigma_1} e^{-\dfrac{(f_i-\bar{F}_1)^2}{2\sigma_1^2}} + \frac{1-c}{\sigma_2} e^{-\dfrac{(f_i-\bar{F}_2)^2}{2\sigma_2^2}}\right] = \\ &= \mathcal{L}^{(0)} + \mathcal{L}^{(1)}(f_1, f_2, \ldots, f_N; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c).\end{aligned} \tag{15}$$

Точечные оценки параметров $\bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c$ находятся минимизацией:

$$\mathcal{L} = \mathcal{L}^{(0)} + \mathcal{L}^{(1)}(f_1, f_2, \ldots, f_N; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c) \to \min \Longrightarrow \tag{16}$$

$$\boxed{\mathcal{L}^{(1)}(\mathbf{f}; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c) \to \min}, \qquad \sigma_1, \sigma_2 > 0, \;\; 0 < c < 1. \tag{17}$$

3. Численная минимизация целевой функции

Варианты.

1. Перебор сетки значений многомерной области параметров (может быть эффективен, если $M \leqslant 2$).

2. Поиск методом градиентного спуска.

3. Численный поиск минимума при помощи программы для произвольной функции. В принципе есть программа `e04cgf` пакета NAG (формально она на Fortran 77, но идет и при более поздних версиях транслятора). Но можно применить и любую другую программу на другом языке.

4. Оценка доверительных интервалов

Целевая функция:

$$\mathcal{L}^{(1)}(f_1, f_2, \ldots, f_N; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c) = \mathcal{L}^{(1)}(\mathbf{f}, \mathbf{a}). \tag{18}$$

Примем обозначения:

$$\mathcal{L}_0 \equiv \min \mathcal{L}^{(1)}(\mathbf{f}, \mathbf{a}), \tag{19}$$

$$\mathcal{L}_\mathrm{p}(a_m) \equiv \min_{a_m=\mathrm{const}} \mathcal{L}^{(1)}(\mathbf{f}, \mathbf{a}). \tag{20}$$

Последнюю функцию можно назвать *профилем целевой функции для параметра* a_m. Тогда корни уравнения

$$\boxed{\mathcal{L}_\mathrm{p}(a_m) = \mathcal{L}_0 + 1/2} \tag{21}$$

дают границы доверительных интервалов для оценки параметра a_m на доверительном уровне 1σ. Запись:

$$a_{m,0}{}^{+\sigma_m^+}_{-\sigma_m^-}, \tag{22}$$

где $a_{m,0}$ — точечная оценка, а доверительные полуинтервалы

$$\sigma_m^+ = a_{m,2} - a_{m,0}, \qquad \sigma_m^- = a_{m,0} - a_{m,2}. \tag{23}$$

См. рисунок 1.

Литература: *Худсон Д.* Статистика для физиков. М.: Мир, 1970. 296 с.

5. Визуализация

Нужно сопоставить модельную функцию $\varphi_\mathrm{mod}(f; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c)$ с гистограммой наблюдаемого распределения f на одном рисунке.

Гистограмма наблюдаемого распределения. J ячеек длиной $\Delta f = 0.1\,\mathrm{dex}$; в каждой ячейке N_j объектов, $j = 1, 2, \ldots, J$. $\sum N_j = N$ (при нормировке столбцов гистограммы на N). Пусть ячейки гистограммы включают правые границы (...`--]--]--]--`...). Например, $(-2.40, -2.30]; (-2.30, -2.20], \ldots$

Модельная функция $\varphi_\mathrm{mod}(f)$. Нужно согласовать нормировки гистограммы $\varphi_\mathrm{hist}(f)$ и модельной функции $\varphi_\mathrm{mod}(f)$:

$$\int_{-\infty}^{+\infty} \varphi_\mathrm{hist}(f)\, d\varphi = \sum_{j=1}^{J} N_j\, \Delta f = N\, \Delta f, \tag{24}$$

$$\int_{-\infty}^{+\infty} \varphi_\mathrm{mod}(f)\, d\varphi = 1. \tag{25}$$

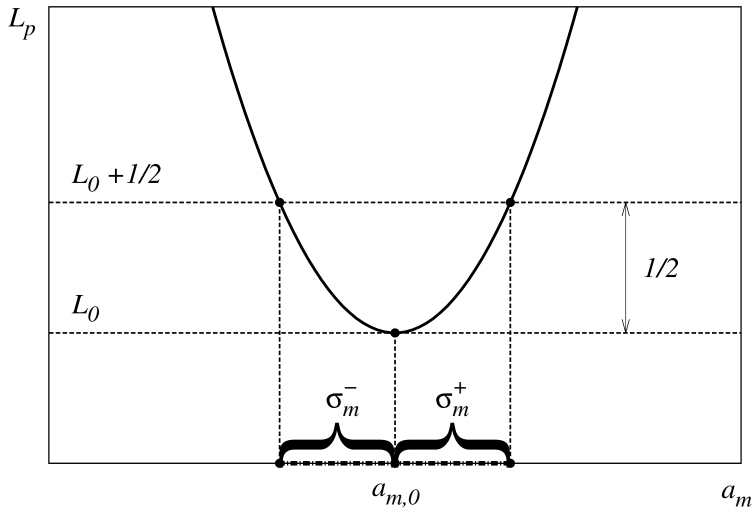

Рис. 1. Профиль $\mathcal{L}_\mathrm{p}(a_m)$ целевой функции $\mathcal{L}^{(1)}(\mathbf{f}, \mathbf{a})$ для параметра a_m и доверительный интервал на уровне 1σ для точечной оценки $a_{m,0}$ параметра.

Т.е. совместно с гистограммой вместо модельной функции $\varphi_\mathrm{mod}(f)$ нужно рисовать

$$\boxed{N\,\Delta f\,\varphi_\mathrm{mod}(f).} \tag{26}$$

Используя формулу (26), изобразить φ_mod и две ее нормальные составляющие [см. (13)].

6. Задание и данные

Для заданного набора $f_i, i = 1, \ldots, N$ получить следующие результаты.

1. Точечные оценки и доверительные интервалы для всех пяти параметров

$$a_{m,0}{}^{+\sigma_m^+}_{-\sigma_m^-}, \qquad m = 1, \ldots, M.$$

2. График $\mathcal{L}_1(a_m)$ с отмеченными на них доверительными интервалами для каждого параметра.

3. График сравнения наблюдаемого распределения с модельным.

Данные: файл `all.dat`.

Указания. В качестве начального приближения можно взять $\bar{F}_1 = -1.5$, $\sigma_1 = 0.5$, $\bar{F}_2 = -0.5$, $\sigma_1 = 0.2$, $c = 0.5$.

Существует опасность ухода за пределы области определения в случаях параметров σ_1, σ_2 и c. Можно применить следующие приемы.

1. Логарифмическое преобразование: например, $\sigma_1 \longleftrightarrow \lg\sigma_1$.

```
...
X(2)=LOG(0.5D0)
X(4)=LOG(0.2D0)
...

SUBROUTINE FUNCT1 (M,X,FC)
SIG1=EXP((X(2))
SIG2=EXP((X(4))
```

2. Тангенциальное преобразование: $c \longleftrightarrow \operatorname{tg}(\pi c - \pi/2)$.