ВВЕДЕНИЕ В МАТЕМАТИЧЕСКУЮ СТАТИСТИКУ

Конспект лекций

Голубев Г. К.

Оглавление

Введение

Если говорить очень кратко о том, какими задачами занимается статистика, то можно в самом общем виде сказать, что это задачи резюмирования данных. Проще всего пояснить, что такое резюмирование на следующем простом примере. Представим себе, что мы собрали данные о зарплатах в городе N. Это будет большое количество чисел и обычный человек скорее всего предпочел бы иметь вместо него одно или несколько чисел, которые описывали бы хорошо все множество зарплат. Эти несколько чисел и представляют собой резюме данных с наивной точки зрения. Обычно в качестве такого числа мы привыкли видеть среднюю зарплату. Вопрос, который при этом мы даже себе часто не задаем состоит в том, а действительно ли средняя заплата хорошо резюмирует все множество зарплат и нет чего-либо лучше этой величины? Если говорить очень грубо, то математическая статистика и пытается ответить на подобного рода вопросы.

Словосочетание *математическая статистика* в названии этих лекций в некоторой степени избавляет меня от необходимости приводить массу примеров, зачастую анекдотических и не имеющих особой реальной ценности. Здесь мы будем говорить в основном о математических обоснованиях статистических методов и верить, что используемые аргументы верны как для статистической обработки зарплат, так например, и для результатов физических измерений фундаментальных физических постоянных таких, как скорость света.

В основе всех наших дальнейших рассуждений лежит гипотеза о том, данные, которыми мы располагаем имеют случайный характер, т.е. представляют собой некоторое множество случайных величин. Хочу подчеркнуть то, что принципиально важно, что мы имеем в своем распоряжении много случайных величин. Дело в том, что свойство случайности проявляется только тогда когда у нас имеется достаточно много объектов. Делать какие либо выводы на основе наблюдения одной случайной величины довольно бессмысленное занятие. Гипотеза о том, что реаль-

ные данные являются случайными конечно подразумевает, что мы хорошо понимаем математические свойства этих объектов. К сожалению, это довольно оптимистический взгляд на реальную ситуацию. Собственно в этом и заключается основная сложность в изучении и использовании математической статистики. Совершенно недостаточно прочитать и выучить определение случайной величины и несколько теорем, чтобы овладеть этим предметом. Нужна довольно существенная практика работы со случайными объектами для того, чтобы выработать определенное понимание и автоматизм в их использовании. Собственно это относится к любой деятельности, которой занимаются люди. Мы все хорошо знаем, что для того чтобы сносно рисовать, недостаточно читать книжки по живописи. Статистика не исключение, чтобы ей пользоваться со здравым смыслом, надо решать много конкретных задач.

Глава 1

Статистические модели

Одной из самых простых со статистической точки зрения является задача резюмирования данных, получаемых в результате физических экспериментов при измерении фундаментальных физических постоянных, например таких, как скорость света. Сами по себе такого рода измерения как правило являются дорогостоящими и в известной степени изощренными. Однако их результатом являются данные, имеющие простую вероятностную структуру. Рассмотрим такой гипотетический эксперимент. Как правило, он проводится много раз и в разных условиях. Предположим, что мы получили n чисел $Y^n = \{Y_1, \ldots, Y_n\}$ в качестве результатов измерений некоторой физической величины, которую будем обозначать для определенности μ. Кроме нас эти данные скорее всего мало кому интересны. Интерес представляет два числа — значение μ и точность, с которой ее удалось померить. Мы можем резюмировать данные Y^n, которые были получены, различными способами. Например,

- средним значением

$$\bar{Y}^n = \arg\min_m \sum_{k=1}^{n} (Y_k - m)^2 = \frac{1}{n} \sum_{k=1}^{n} Y_k$$

- медианой

$$\mathrm{med}(Y^n) = \arg\min_m \sum_{k=1}^{n} |Y_k - m|$$

- или минимальным размахом

$$\arg\min_m \bigl\{ \max_k |Y_k - m| \bigr\} = \frac{\min(Y^n) + \max(Y^n)}{2}$$

Можно придумать еще множество способов это сделать. Однако наша цель — найти самый лучший метод. Получить осмысленное решение этой задачи это возможно только в рамках *вероятностной модели* для данных Y^n. В рассматриваемом случае естественная с физической точки зрения и самая простая в математическом смысле модель имеет вид

$$Y_k = \mu + \epsilon_k, \quad k = 1, \dots, n, \tag{1.0.1}$$

где

- ϵ_k — независимые, одинаково распределенные случайные величины с нулевым средним,
- $\mu \in \mathbb{R}$ — неизвестный параметр, который мы хотим оценить по наблюдениям Y^n.

Оценить параметр μ означает найти такую функцию $\hat{\mu}(Y_1, \dots, Y_n) : \mathbb{R}^n \to \mathbb{R}$, которая была бы как можно ближе к μ.

1.1 Функция распределения и плотность

С вероятностной точки зрения модель (1.0.1) эквивалентна предположению о том, что распределение наблюдений Y^n полностью определяется следующей функцией:

$$P(y_1, \dots, y_n; \mu) = \mathbf{P}\{Y_1 \le y_1, \dots, Y_n \le y_n\} = \prod_{k=1}^{n} F(y_k - \mu);$$

здесь

$$F(x) = \mathbf{P}\{\epsilon_k \le x\}$$

функция распределения случайной величины ϵ_k. Символ ; разделяет аргументы функции на собственно аргументы и параметры, т.е. величины, которые считаются фиксированными.

Простейшие свойства функции распределения

- $0 \le F(x) \le 1$;
- $F(x)$ — неубывающая функция;
- $\lim\limits_{x \to -\infty} F(x) = 0$ и $\lim\limits_{x \to +\infty} F(x) = 1$.

Производная функции распределения (если она существует)

$$p(x) = \frac{dF(x)}{dx}$$

называется *(вероятностной) плотностью распределения* случайной величины ϵ_k. Простейшие свойства вероятностной плотности

- $p(x) \geq 0$ при всех $x \in \mathbb{R}$;
- $\int_{-\infty}^{\infty} p(x)\,dx = 1$.

Физический смысл плотности очень прозрачен:

$$\mathbf{P}\big\{\epsilon_k \in [x, x+h]\big\} \approx p(x) \cdot h$$

при малых h. То есть вероятность того, что случайная величина ϵ_k принадлежит маленькому отрезку $[x, x+h]$ пропорциональна длине отрезка h с коэффициентом пропорциональности равным плотности. Это свойство можно выразить в чуть более общей форме. Пусть $B_h(x)$ — некоторое измеримое множество такое, что

$$\max_{y \in B_h(x)} |x - y| \leq h.$$

Тогда при малых h

$$\mathbf{P}\big\{\epsilon_k \in B_h(x)\big\} \approx p(x) \times \int_{B_h(x)} dx = p(x) \times \{B_h(x)\};$$

здесь $\mathrm{mes}\{B_h(x)\}$ — мера Лебега множества $B_h(x)$.

Совершенно аналогично плотность определяется и для многомерных случайных величин. Например,

$$p(y_1, y_2) = \frac{\partial}{\partial y_2}\frac{\partial}{\partial y_1}\mathbf{P}\big\{\epsilon_1 \leq y_1, \epsilon_2 \leq y_2\big\}.$$

Свойства многомерных плотностей аналогичны свойствам одномерных. Но у многомерных плотностей есть исключительно важное свойство, а именно, если случайные величины $\epsilon_1, \ldots, \epsilon_n$ независимы, то

$$p(y_1, \ldots, y_n) = \prod_{k=1}^{n} p(y_k).$$

1.2 Несколько полезных фактов о функции распределения

Неравенство Чернова

В русскоязычной литературе это неравенство называют экспоненциальным неравенством Чебышева.

Теорема 1.2.1

$$1 - F(x) \le \inf_{\lambda>0}\left\{ \mathrm{e}^{-\lambda x}\mathbf{E}\mathrm{e}^{\lambda\epsilon}\right\};$$

здесь

$$\mathbf{E}\mathrm{e}^{\lambda\epsilon} = \int_{-\infty}^{\infty} \mathrm{e}^{\lambda x} p^{\epsilon}(x)\,dx = \int_{-\infty}^{\infty} \mathrm{e}^{\lambda x}\,dF(x).$$

Доказательство. Заметим, что для любого положительного числа $\lambda > 0$ справедливо неравенство

$$\mathbf{1}\{\epsilon > x\} \le \mathrm{e}^{\lambda(\epsilon - x)}.$$

Отсюда сразу же получаем

$$1 - F_{\epsilon}(x) = \mathbf{E}\mathbf{1}\{\epsilon > x\} \le \mathbf{E}\mathrm{e}^{\lambda(\epsilon-x)} = \mathrm{e}^{-\lambda x}\mathbf{E}\mathrm{e}^{\lambda\epsilon}.$$

Поскольку это неравенство справедливо при всех положительных λ, то оно станет только лучше, если мы возьмем минимум по $\lambda > 0$. ■

Ценность того или иного результата определяется не столько сложностью его доказательства а тем как часто он используется. В этом смысле неравенство Чернова, по-видимому, рекордный результат. Объясняется это прежде всего тем, что очень часто в статистике нас интересуют вероятности больших уклонений сумм независимых случайных величин

$$\mathbf{P}\left\{\sum_{i=1}^{n}\epsilon_i \ge x\right\}.$$

Для оценки этой величины неравенство Чернова является исключительно удобным инструментом поскольку

$$\mathbf{E}\exp\left\{\lambda\sum_{i=1}^{n}\epsilon_i\right\} = \prod_{i=1}^{n}\mathbf{E}\exp(\lambda\epsilon_i).$$

С помощью этого тождества и неравенства Чернова контроль вероятностей больший уклонений становится, как правило, рутинной математической работой. При этом важно, что это неравенство правильно описывает поведение

$$\log\left\{\mathbf{P}\left[\sum_{i=1}^{n}\epsilon_i \geq x\right]\right\}$$

при больших x.

Заметим, что если вместо функции $\exp(x)$ в неравенстве

$$\mathbf{1}\{\xi \geq x\} \leq \exp[\lambda(\xi - x)]$$

мы использовали бы степенную функцию $|x|^m$, то получили бы неравенство

$$\mathbf{1}\{\xi \geq x\} \leq \frac{|\xi|^m}{x^m}$$

и как следствие стандартное неравенство Чебышева

$$\mathbf{P}\big\{\xi \leq x\big\} \leq \frac{\mathbf{E}|\xi|^m}{x^m}.$$

Заметим однако, что вычислять при больших p величину $\mathbf{E}\left|\sum_{i=1}^{n}\epsilon_i\right|^p$ неизмеримо сложнее, чем $\mathbf{E}\exp(\lambda\epsilon_i)$.

Метод инверсии

Второй исключительно простой, но полезный факт состоит в следующем. Предположим, что функция распределения случайной величины ϵ

$$F(x) = \mathbf{P}\big\{\epsilon \leq x\big\}$$

имеет обратную функцию $F^{-1}(y)$, $y \in (0,1)$, которая непрерывна. То есть, для любого $x \in \mathbb{R}$

$$F^{-1}[F(x)] = x$$

Пусть U — случайная величина равномерно распределенная на отрезке $[0,1]$

$$\mathbf{P}\big\{U \leq x\big\} = \begin{cases} 0, & x < 0, \\ x, & x \in [0,1], \\ 1, & x > 1. \end{cases}$$

Тогда

$$\mathbf{P}\Big\{F^{-1}(U) \leq x\Big\} = \mathbf{P}\Big\{F[F^{-1}(U)] \leq F(x)\Big\} = \mathbf{P}\Big\{U \leq F(x)\Big\} = F(x).$$

Это значит, что случайная величина $F^{-1}(U)$ имеет функцию распределения $F(x)$. То есть для того, чтобы генерировать случайную величину с функцией распределения $F(x)$ достаточно уметь генерировать равномерно распределенную случайную величину и уметь обращать $F(x)$. Первая задача является безусловно очень сложной и в этом курсе мы ее обсуждать не будем, считая, что у нас имеется достаточно хороший генератор независимых, равномерно распределенных на отрезке $[0,1]$ случайных величин.

Вторая задача в общем случае тоже не тривиальна, но в некоторых случаях имеет простые решения.

Пусть случайная величина ζ имеет *экспоненциальное распределение* с параметром λ. Это значит, что ее функция распределения имеет следующий вид

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - \exp(-x/\lambda), & x \geq 0. \end{cases}$$

Отсюда понятно, что

$$F^{-1}(u) = -\lambda \log(1-u),\ u \in (0,1).$$

Заметим, что если U равномерно распределена на $[0,1]$, то и $1-U$ обладает этим же свойством. Таким образом, для того, чтобы получить случайную величину ζ с экспоненциальным распределением, достаточно взять

$$\zeta = -\lambda \log(U).$$

Столь же просто генерируются случайные величины с *распределением Лапласа*, плотность распределения которого имеет вид

$$p(x; m, \lambda) = \frac{1}{2\lambda} \exp\left\{ -\frac{|x-m|}{\lambda} \right\}$$

Величина $m \in \mathbb{R}$ называется *параметром сдвига*, распределения Лапласа, а $\lambda \in \mathbb{R}^+$ — *параметром масштаба*.

Метод Бокса-Мюллера.

Генерирование случайных величин со *стандартным гауссовским распределением* уже не столь простая задача. Функция распределения такой величины имеет вид

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-u^2/2}\, du.$$

В принципе, для генерирования таких величин можно было бы воспользоваться методом, обратив эту функцию распределения, но мы рассмотрим более оригинальный метод.

Возьмем пару независимых стандартных гауссовских случайных величин ξ_1, ξ_2. Плотность ее распределения имеет очевидно следующий вид:

$$p(x,y) = \frac{1}{2\pi}\exp\left\{-\frac{x^2+y^2}{2}\right\}.$$

Попробуем теперь найти пару случайных величин r и ϕ такую, чтобы случайные величины

$$\begin{aligned}\xi_1' =& r\cos(\phi),\\ \xi_2' =& r\sin(\phi)\end{aligned}$$

имели бы совместную плотность $p_{\xi_1\xi_2}(x,y)$. Обозначим плотность распределения ξ_1', ξ_2' как $q(r,\phi)$.

Рассмотрим маленький элемент на плоскости

$$\Delta(r,\phi) = \{r',\phi' : r' \in [r, r+\Delta_r],\ \phi' \in [\phi, \phi+\Delta_\phi]\}.$$

Вероятность того что пара r,ϕ принадлежит этому элементу равна

$$q(r,\phi)\Delta_r\Delta_\phi.$$

Но поскольку площадь этого элемента равна $r\Delta_\phi\Delta_r$, то вероятность того, что гауссовская пара попадет в него равна

$$\frac{1}{2\pi}\exp(-r^2/2)r\Delta_r\Delta_\phi.$$

Поэтому

$$q(r,\phi) = \frac{1}{2\pi}r\exp(-r^2/2).$$

Отсюда видим, что r и ϕ — независимые случайные величины, причем ϕ равномерно распределена на $[0,2\pi]$, а r имеет плотность распределения

$$q_r(x) = x\exp(-x^2/2).$$

Это означает, что функция распределения этой случайной величины вычисляется как

$$F_r(x) = 1-\exp(-x^2/2)$$

и ее обратная функция вычисляется очень просто

$$F_r^{-1}(u) = \sqrt{-2\log(1-u)}.$$

Суммируя выкладки, приходим к следующему методу генерирования пары независимых гауссовских случайных величин ξ_1, ξ_2 из пары независимых равномерно распределенных случайных величин U_1, U_2

$$\begin{aligned}\xi_1 &= \sqrt{-2\log(U_1)}\cos(2\pi U_1),\\ \xi_2 &= \sqrt{-2\log(U_1)}\sin(2\pi U_2).\end{aligned}$$

1.3 Ансамбли случайных величин

1.3.1 Эмпирическая функция распределения

В статистике для вероятностных объектов обычно есть их эмпирические аналоги, которые со статистической точки зрения являются оценками этих объектов. Для функции распределения случайной величины ϵ $F_\epsilon(x) = \mathbf{P}\{\epsilon \le x\}$ ее эмпирическим (т.е. построенным по данным $\epsilon^n = \{\epsilon_1, \ldots, \epsilon_n\}$) аналогом является *эмпирическая функция распределения*

$$\bar{F}(x; \epsilon^n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\epsilon_i \le x\}.$$

Очевидно, что эта функция не убывает и не меняет своего значения при любых перестановках величин $\{\epsilon_i,\ i = 1, \ldots, n\}$ и поэтому

$$\bar{F}(x; \epsilon^n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\epsilon_{(i)} \le x\},$$

где

$$\epsilon_{(1)} \le \epsilon_{(2)} \le \cdots \le \epsilon_{(n)}$$

— упорядоченные в неубывающем порядке значения величин $\{\epsilon_i,\ i = 1, \ldots, n\}$. Величины $\epsilon_{(i)}$ называются *порядковыми статистиками* .

Эмпирическая функция распределения по переменной x является кусочно-постоянной и имеет скачки величины $1/n$ в точках $\epsilon_{(i)},\ i = 1, \ldots, n$.

Смысл введения эмпирических аналогов состоит единственно в том, что они должны быть близки к своим прототипам. Давайте попробуем

это проверить предполагая, что ϵ_i — *независимые, одинаково распределенные* случайные величины с функцией распределения $F(x)$. Множество $\epsilon^n = \{\epsilon_1, \ldots, \epsilon_n\}$ таких случайных величин называют *повторной выборкой.*

Тогда при фиксированном x мы находим

$$\mathbf{E}\bar{F}(x;\epsilon^n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\mathbf{1}\{\epsilon_i \le x\} = \mathbf{P}\{\epsilon_1 \le x\} = F(x)$$

и

$$\begin{aligned}\mathbf{E}[\bar{F}(x;\epsilon^n) - F(x)]^2 =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\mathbf{E}\big[\mathbf{1}\{\epsilon_i \le x\} - F(x)\big]\big[\mathbf{1}\{\epsilon_k \le x\} - F(x)\big] \\ =& \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}\big[\mathbf{1}\{\epsilon_i \le x\} - F(x)\big]^2 \\ =& \frac{1}{n^2}\sum_{i=1}^{n}\Big[F(x) - 2F^2(x) + F^2(x)\Big] \\ =& \frac{F(x)[1-F(x)]}{n}.\end{aligned}$$

Отсюда, в частности, в силу центральной предельной теоремы получим

Теорема 1.3.1

$$\lim_{n\to\infty}\mathbf{P}\left\{\frac{\sqrt{n}\big|\bar{F}(x;\epsilon^n) - F(x)\big|}{\sqrt{F(x)[1-F(x)]}} \ge z\right\} = \frac{2}{\sqrt{2\pi}}\int_z^\infty \mathrm{e}^{-u^2/2}\,du.$$

Таким образом, видим, что при фиксированном x эмпирическая функция распределения отличается от функции распределения на величину порядка $1/\sqrt{n}$.

Естественно, возникает вопрос насколько близки эти объекты как функции. Чтобы ответить на него, определим два гауссовских процесса.

Определение 1 *Гауссовский случайный процесс $W(t),\ t \ge 0$ с нулевым средним и корреляционной функцией*

$$\mathbf{E}W(t)W(s) = \min\{t, s\}$$

называется винеровским процессом.

Определение 2 *Случайный процесс*

$$W_\circ(t) = W(t) - tW(1),\ t \in [0,1]$$

называется броуновским мостом.

Частично ответ на вопрос о близости функции распределения и ее эмпирической версии дает

Теорема 1.3.2 (Колмогоров) *Если $F(\cdot)$ имеет непрерывную обратную функцию*

$$\lim_{n\to\infty} \sqrt{n}\big[\bar{F}(x;\epsilon^n) - F(x)\big] \stackrel{\mathbf{P}}{=} W_\circ[F^{-1}(x)].$$

Расстояние между функциями $\bar{F}(x;\epsilon^n)$ и $F_\epsilon(x)$ можно измерять величиной

$$\sup_x \big|\bar{F}(x;\epsilon^n) - F(x)\big| \stackrel{\text{def}}{=} \|\bar{F}(\cdot;\epsilon^n) - F\|_\infty,$$

которая называется *расстоянием Колмогорова*. Поэтому далее нас будет интересовать распределение величины $\|\bar{F}(\cdot;\epsilon^n) - F\|_\infty$, а именно,

$$\mathbf{P}\Big\{\|\bar{F}(\cdot;\epsilon^n) - F\|_\infty > z\Big\}.$$

Прежде чем эту величину вычислить, докажем очень простой факт о том, что если $F(x)$ обратима, то распределение $\|\bar{F}(\cdot;\epsilon^n) - F\|_\infty$ не зависит от F. Точнее, справедлив следующий результат.

Теорема 1.3.3 *Пусть $U_1, \ldots, U_n$ — независимые случайные величины равномерно распределенные на отрезке $[0,1]$. Если обратная функция $F^{-1}(y)$, $y \in (0,1)$ существует и непрерывна, то*

$$\mathbf{P}\Big\{\|\bar{F}_\epsilon(\cdot;\epsilon^n) - F\|_\infty > z\Big\} = \mathbf{P}\Big\{\sup_{u\in[0,1]} |\bar{F}(u;U^n) - u| > z\Big\};$$

здесь

$$\bar{F}(u;U^n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{U_i \le u\}.$$

Доказательство. Этот результат вытекает, по-существу, из метода инверсии. Действительно, если обратная функция $F^{-1}(y)$, $y \in (0,1)$ существует и непрерывна, то

$$\begin{aligned}\mathbf{P}\Big\{\|\bar{F}(\cdot;\epsilon^n) - F\|_\infty > z\Big\} &= \mathbf{P}\Big\{\sup_{u\in[0,1]}\big|\bar{F}[F^{-1}(u);\epsilon^n] - F[F^{-1}(u)]\big| > z\Big\} \\ &= \mathbf{P}\Big\{\sup_{u\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}\mathbf{1}\{\epsilon_i \le F^{-1}(u)\} - F(F^{-1}(u))\Big| > z\Big\} \\ &= \mathbf{P}\Big\{\sup_{u\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}\mathbf{1}\{F^{-1}(U_i) \le F^{-1}(u)\} - u\Big| > z\Big\} \\ &= \mathbf{P}\Big\{\sup_{u\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}\mathbf{1}\{U_i \le u\} - u\Big| > z\Big\}. \quad \blacksquare\end{aligned}$$

Практическое значение этой теоремы заключается в том, что нам не нужно вычислять функцию распределения $\sup_x\big|\bar{F}(x;\epsilon^n) - F_\epsilon(x)\big|$ для всех функций распределения F, это достаточно сделать один раз для равномерного распределения.

Как уже отмечалось, эмпирическая функция равномерно распределенных на отрезке $[0,1]$ случайных величин $U_1,\dots,U_n$

$$\bar{F}(u;U^n) = \frac{1}{n}\sum_{i=1}\mathbf{1}\{U_{(i)} \le u\}$$

полностью определяется порядковыми статистиками

$$U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}.$$

Распределение этих порядковых статистик вычисляется на основании следующего результата:

Теорема 1.3.4 (Pyke (1965))

$$U_{(i)} \stackrel{\mathbf{P}}{=} \sum_{s=1}^{i}\zeta_s \Big/ \sum_{s=1}^{n+1}\zeta_s,$$

где $\zeta_s,\ s = 1,\dots,n+1$ *— независимые, стандартные, экспоненциально распределенные случайные величины*

$$\mathbf{P}\big\{\zeta_s \le z\big\} = 1 - \exp(-z),\ z \ge 0.$$

Этим результатом мы будем пользоваться активно далее. Сейчас же заметим, что

$$\begin{aligned}
&\mathbf{P}\Big\{\sup_{u\in[0,1]}\sqrt{n}|\bar{F}(u;U^n)-u|\geq z\Big\}\\
&=\mathbf{P}\Big\{\sup_{k=1,\dots,n}\sqrt{n}|U_{(k)}-k/n|\geq z\Big\}\\
&=\mathbf{P}\left\{\sup_{k=1,\dots,n}\left|n\sum_{s=1}^{k}\zeta_s-k\sum_{s=1}^{n+1}\zeta_s\right|\geq z\sqrt{n}\sum_{s=1}^{n+1}\zeta_s\right\}\\
&=\mathbf{P}\left\{\sup_{k=1,\dots,n}\left|\frac{1}{\sqrt{n}}\sum_{s=1}^{k}[\zeta_s-1]-\frac{k}{n\sqrt{n}}\sum_{s=1}^{n+1}[\zeta_s-1]-\frac{1}{n}\right|\geq\frac{z}{n}\sum_{s=1}^{n+1}\zeta_s\right\}.
\end{aligned}\tag{1.3.2}$$

Рассмотрим случайный процесс

$$W_n(t)=\frac{1}{\sqrt{n}}\sum_{s=1}^{\lfloor nt\rfloor}[\zeta_s-1]$$

Очевидно, что $\mathbf{E}W_n(t)=0$ и при $n\to\infty$

$$\begin{aligned}
\mathbf{E}W_n(t)W_n(s)&=\frac{1}{n}\sum_{i=1}^{\min\{\lfloor nt\rfloor,\lfloor ns\rfloor\}}\mathbf{E}[\zeta_s-1]^2\\
&=\frac{\min\{\lfloor nt\rfloor,\lfloor ns\rfloor\}}{n}=\min\{t,s\}+O\left(\frac{1}{n}\right).
\end{aligned}\tag{1.3.3}$$

Из (1.3.3) и центральной предельной теоремы вытекает, что при $n\to\infty$

$$W_n(t)\xrightarrow{\mathbf{P}}W(t)$$

и

$$\frac{1}{\sqrt{n}}\sum_{s=1}^{\lfloor nt\rfloor}[\zeta_s-1]-t\frac{1}{\sqrt{n}}\sum_{s=1}^{n}[\zeta_s-1]\xrightarrow{\mathbf{P}}W(t)-tW(1)=W_\circ(t).$$

Под сходимостью в этих формулах имеется ввиду сходимость любых конечномерных распределений.

Поэтому из (1.3.2) практически очевидно, что

$$\lim_{n\to\infty}\mathbf{P}\Big\{\sup_{u\in[0,1]}\sqrt{n}|F(u;U^n)-u|\geq z\Big\}=\mathbf{P}\Big\{\sup_{t\in[0,1]}|W_\circ(t)|\geq z\Big\}.$$

Этот факт и теорема 1.3.3 лежат в основе доказательства следующего результата:

Теорема 1.3.5 (Колмогоров)

$$\lim_{n\to\infty} \mathbf{P}\Big\{\sqrt{n}\|\bar{F}(\cdot;\epsilon^n) - F\|_\infty > z\Big\} = 2\sum_{k=1}^{\infty}(-1)^{k-1}\mathrm{e}^{-2k^2z^2}.$$

Неасимптотическая форма этой теоремы следующий имеет вид:

Теорема 1.3.6 (Неравенство Дворецкого-Кифера-Вольфовица)

$$\mathbf{P}\Big\{\sqrt{n}\|\bar{F}(\cdot;\epsilon^n) - F\|_\infty > z\Big\} \le 2\mathrm{e}^{-2z^2}.$$

Доказательство этого удивительно точного неасимптотического результата совсем нетривиально и было получено сравнительно недавно.

Из этих результатов вытекает важный вывод, что расстояние между функцией распределения и ее эмпирическим аналогом в равномерной норме имеет тот же порядок $1/\sqrt{n}$, что и расстояние в фиксированной точке.

1.3.2 Экстремальные значения

Теорема Пайка и метод инверсии позволяют довольно просто находить предельные распределения экстремальных значений выборок из независимых случайных величин. Другими словами позволяют понять как устроен с вероятностной точки зрения носитель эмпирической функции распределения.

Гауссовское распределение

Пусть $\epsilon_i,\ i = 1,\dots,n$ — *стандартные, независимые, гауссовские* случайные величины. Нас будет интересовать распределение максимума этих случайных величин, т.е.

$$\mathbf{P}\Big\{\max_{i=1,\dots,n}\epsilon_i \le x\Big\}.$$

Согласно методу инверсии

$$\epsilon_{(n)} \overset{\mathbf{P}}{=} F^{-1}[U_{(n)}] = F^{-1}[1-(1-U_{(n)})],$$

где $U_1,\dots,U_n$ — независимые, равномерно распределенные на $[0,1]$ случайные величины, $F^{-1}(y)$ — обратная функция для гауссовской функции распределения

$$F(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-u^2/2}\,du = 1 - \frac{1}{\sqrt{2\pi}}\int_{x}^{\infty}\mathrm{e}^{-u^2/2}\,du.$$

Из теоремы Pyke'a находим

$$U_{(n)} = \sum_{i=1}^{n} \zeta_i \Big/ \sum_{i=1}^{n+1} \zeta_i = 1 - \zeta_n \Big/ \sum_{i=1}^{n+1} \zeta_i,$$

где ζ_i — независимые, стандартные экспоненциально распределенные случайные величины. В силу центральной предельной теоремы при $n \to \infty$

$$\begin{aligned} U_{(n)} =& 1 - \frac{\zeta_{n+1}}{n+1}\left[1 + \frac{1}{n+1}\sum_{i=1}^{n+1}(\zeta_i - 1)\right]^{-1} \\ =& 1 - \frac{\zeta_{n+1}}{n+1}\left[1 + \frac{1}{\sqrt{n+1}} \cdot \frac{1}{\sqrt{n+1}}\sum_{i=1}^{n+1}(\zeta_i - 1)\right]^{-1} \\ =& 1 - \frac{\zeta_{n+1}}{n+1}\left[1 + O\left(\frac{1}{\sqrt{n+1}}\right)\right]. \end{aligned} \tag{1.3.4}$$

То есть величина $1 - U_{(n)}$ при больших n мала и имеет порядок $1/n$. Поэтому функцию гауссовского распределения $F(x)$ нужно обращать при $x \to \infty$. Интегрируя по частям, находим что, при $x \to \infty$

$$\begin{aligned} F(x) =& 1 + \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{1}{u}\, d\mathrm{e}^{-u^2/2} \\ =& 1 - \frac{\exp(-x^2/2)}{x\sqrt{2\pi}} - \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{\mathrm{e}^{-u^2/2}}{u^2}\, du \\ =& 1 - \frac{\exp(-x^2/2)}{x\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{1}{u^3}\, d\mathrm{e}^{-u^2/2} \\ =& 1 - \frac{\exp(-x^2/2)}{\sqrt{2\pi}}\left[\frac{1}{x} + O\left(\frac{1}{x^3}\right)\right]. \end{aligned}$$

Поэтому при $y \to 1$ для $z = F^{-1}(y)$ получаем уравнение

$$1 - \frac{\exp(-z^2/2)}{\sqrt{2\pi}}\left[\frac{1}{z} + O\left(\frac{1}{z^3}\right)\right] = y$$

или, что эквивалентно

$$\frac{z^2}{2} - \log\left[\frac{1}{z} + O\left(\frac{1}{z^3}\right)\right] = \log\frac{1}{(1-y)\sqrt{2\pi}}.$$

То есть величина z является корнем уравнения

$$z = L_y(z)$$

где

$$L_y(z) = \left\{ 2\log \frac{1}{(1-y)\sqrt{2\pi}} + 2\log\left[\frac{1}{z} + O\left(\frac{1}{z^3}\right)\right]\right\}^{1/2}.$$

Функция $L_y(z)$ медленно меняется (убывает) при больших z. Ее производная ограничена снизу величиной

$$L'_y(z) > -\frac{1}{2z}\left\{ 2\log \frac{1}{(1-y)\sqrt{2\pi}} + 2\log\left[\frac{1}{z} + O\left(\frac{1}{z^3}\right)\right]\right\}^{-1/2}.$$

Сама же функция ограничена сверху постоянной величиной

$$L_y(z) \le \left\{ 2\log \frac{1}{(1-y)\sqrt{2\pi}}\right\}^{1/2}.$$

Отсюда понятно, что величина z находится недалеко от точки

$$z_0 = \left\{ 2\log \frac{1}{(1-y)\sqrt{2\pi}}\right\}^{1/2},$$

которую можно рассматривать как начальную точку для решения уравнения $z = L_y(z)$. Отсюда получаем

$$\begin{aligned} z =& F_\epsilon^{-1}(y) = L_y(z_0) \\ =& \left[2\log \frac{1}{(1-y)\sqrt{2\pi}} - 2\log\log \frac{1}{(1-y)^2 2\pi} + o(1)\right]^{1/2}. \end{aligned} \tag{1.3.5}$$

Таким образом, из (1.3.4) находим

$$\begin{aligned} \epsilon_{(n)} \stackrel{\mathbf{P}}{=}& F_\epsilon^{-1}[U_{(n)}] \\ =& \left[2\log \frac{n+1}{\zeta_n\sqrt{2\pi}} - 2\log\log \frac{(n+1)^2}{\zeta_n^2 2\pi} + o(1)\right]^{1/2} \\ =& \left[2\log \frac{n+1}{\sqrt{8\pi}\log(n+1)} - \log(\zeta_n) + o(1)\right]^{1/2}. \end{aligned} \tag{1.3.6}$$

Отсюда мы приходим к следующему результату.

Теорема 1.3.7 *Пусть $\epsilon^n = \{\epsilon_1, \dots, \epsilon_n\}$ — независимые, стандартные гауссовские случайные величины. Тогда*

$$\lim_{n\to\infty} \mathbf{P}\Big\{ 2m_n\big[\max(\epsilon^n) - m_n\big] \le x\Big\} = \exp[-\exp(-x)],$$

где

$$m_n = \left[2\log \frac{n+1}{\sqrt{8\pi}\log(n+1)}\right]^{1/2}.$$

Доказательство. Из (1.3.6) и формулы Тейлора находим

$$\epsilon_{(n)} = [m_n^2 - \log(\zeta_n) + o(1)]^{1/2} = m_n - \log(\zeta_n)/(2m_n) + o(1).$$

Поэтому

$$\lim_{n\to\infty} \mathbf{P}\Big\{2m_n\big[\max(\epsilon^n) - m_n\big] \le x\Big\} = \mathbf{P}\Big\{-\log(\zeta_n) \le x\Big\}$$
$$= \mathbf{P}\Big\{\zeta_n > \exp(-x)\Big\} = \exp[-\exp(-x)].$$

Таким образом мы видим, что максимум из n стандартных, независимых гауссовских случайных величин концентрируется вблизи величины $m_n \approx \sqrt{2\log(n)}$, причем, концентрация происходит со скоростью $1/m_n$. Эффект концентрации связан с тем, что логарифм "хвоста" стандартного гауссовского распределения

$$\log\left[\frac{1}{\sqrt{2\pi}}\int_x^\infty \mathrm{e}^{-u^2/2}\,du\right]$$

при больших x ведет себя как $-x^2/2$.

Заметим, что если у функции распределения нет такого быстрого приближения к 1 при $x \to \infty$, то и концентрации уже не будет.

Распределение Лапласа

Пусть ϵ_i независимые случайные величины имеющие стандартное распределение Лапласа. В этом случае функция распределения имеет вид

$$F(x) = \begin{cases} 1 - \exp(-x)/2, & x > 0 \\ \exp(x)/2, & x \le 0. \end{cases}$$

Обратная функция распределения вычисляется в данном случае просто

$$F^{-1}(y) = \log\frac{1}{2(1-y)},\ y \in [1/2, 1].$$

Потому из (1.3.4)

$$\max_{i=1,\dots,n} \epsilon_i \stackrel{\mathbf{P}}{=} \log[(n+1)/2] - \log(\zeta_{n+1}) + o(1)$$

и, следовательно, мы доказали следующий результат:

Теорема 1.3.8 *Если $\epsilon_i,\ i = 1, \dots, n$ — независимые случайные величины, имеющие стандартное распределение Лапласа, то*

$$\lim_{n\to\infty} \mathbf{P}\Big\{ \max_{i=1,\dots,n} \epsilon_i - \log(n/2) \le x \Big\} = \exp[-\exp(-x)].$$

Заметим, что хотя по сравнению с гауссовским случаем здесь многое поменялось в поведении максимума, а именно, изменилась скорость, с которой максимум убегает на бесконечность и пропал эффект концентрации, но предельное распределение тем не менее сохранилось.

Распределение Коши

Рассмотрим еще один пример распределения максимума n независимых случайных величин, которые имеют стандартное *распределение Коши* . Плотность этого распределения имеет следующий вид:

$$p_\epsilon(x) = \frac{1}{\pi(1+x^2)}.$$

У этого распределения очевидно нет даже ограниченного абсолютного первого момента

$$\int_{-\infty}^{\infty} |x| p_\epsilon(x)\, dx = \infty.$$

Функция распределения и ее обратная для распределения Коши легко вычисляются

$$F_\epsilon(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x),$$
$$F_\epsilon^{-1}(y) = \tan\left[\pi\left(y - \frac{1}{2}\right)\right].$$

Поэтому при $z \to 0$ из формулы Тейлора найдем

$$F_\epsilon^{-1}(1-z) = \tan\left(\frac{\pi}{2} - \pi z\right) = \frac{1+o(1)}{\pi z}.$$

Из метода инверсии мы получаем

$$\epsilon_{(n)} = F_\epsilon^{-1}(U_{(n)}) = F_\epsilon^{-1}[1-(1-U_{(n)})].$$

Поскольку из теоремы Рукe (см. формулу (1.3.4)) вытекает, что

$$1 - U_{(n)} = \frac{\zeta_{n+1}}{n+1}\left[1 + O\left(\frac{1}{\sqrt{n+1}}\right)\right],$$

то при $n \to \infty$

$$\epsilon_{(n)} = \frac{(1+o(1))n}{\pi\zeta_{n+1}}.$$

Поэтому

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{\pi \max_{i=1,\dots,n} \epsilon_i}{n} \le x\right\} = \mathbf{P}\left\{\frac{1}{\zeta_{n+1}} \le x\right\} = \mathbf{P}\left\{\zeta_{n+1} \ge \frac{1}{x}\right\} = \exp\left(-\frac{1}{x}\right).$$

Таким образом, доказан следующий результат:

Теорема 1.3.9 *Пусть $\epsilon_i,\ i = 1,\dots,n$ — независимые случайные величины, распределенные по стандартному закону Коши. Тогда*

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{\pi}{n} \max_{i=1,\dots,n} \epsilon_i \le x\right\} = \exp\left(-\frac{1}{x}\right).$$

Грубо говоря, эта теорема утверждает, что максимум из n независимых случайных величин, распределенных по стандартному закону Коши имеет порядок n/π. Заметим также, что при больших x

$$\exp\left(-\frac{1}{x}\right) \approx 1 - \frac{1}{x}.$$

Это означает, что это предельное распределение столь же "плохое" , как и распределение Коши. В частности, у него нет первого момента.

1.3.3 Неасимтотические аппроксимации

Завершим этот раздел кратким сравнением предельных теорем об экстремальных значениях и центральной предельной теоремы. На первый взгляд это очень похожие результаты. Центральная предельная теорема утверждает, что если у независимых, одинаково распределенных случайных величин ϵ_i, имеющих нулевое среднее $\mathbf{E}\epsilon_i = 0$ и единичную дисперсию, ограничен абсолютный третий момент $\mathbf{E}|\epsilon_i|^3 < \infty$, то

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i > x\right\} = \frac{1}{\sqrt{2\pi}} \int_x^\infty \mathrm{e}^{-u^2/2}\, du. \tag{1.3.7}$$

Предположим, что мы хотим использовать этот результат в допредельной ситуации т.е. когда число n фиксировано. Заметим, что

$$\begin{aligned}\frac{1}{\sqrt{2\pi}}\int_x^\infty \mathrm{e}^{-u^2/2}\,du &= \frac{1}{\sqrt{2\pi}}\int_0^\infty \mathrm{e}^{-(v+x)^2/2}\,dv \\ &= \frac{1}{\sqrt{2\pi}}\int_0^\infty \mathrm{e}^{-v^2/2-x^2/2-xv}\,dv \\ &\le \frac{1}{\sqrt{2\pi}}\int_0^\infty \mathrm{e}^{-v^2/2-x^2/2}\,dv \le \frac{\mathrm{e}^{-x^2/2}}{2}.\end{aligned}$$

Поэтому (см. (1.3.7)) кажется правдоподобным с наивной точки зрения, что следующее неравенство

$$\mathbf{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i > x\right\} \le \mathrm{e}^{-x^2/2} \tag{1.3.8}$$

справедливо при всех $x > 0$.

В действительности, эта гипотеза в общем случае, конечно, не верна. Причем, неравенство (1.3.8) не верно уже для очень хороших случайных величин, например лапласовских. Оно имеет асимптотический характер. Это означает, что чем больше x, тем больше, вообще говоря, должно быть число наблюдений n чтобы неравенство (1.3.8) было выполнено. То есть, это неравенство выполняется для всех $n \ge n_\circ(x)$. Причем центральная предельная теорема ничего нам не говорит о функции $n_\circ(x)$.

Что касается предельных теорем для экстремальных значений, то они являются более равномерными по x. Например, для лапласовских случайных величин мы можем опускать предел по $n \to \infty$ и использовать предельное равенство

$$\lim_{n\to\infty} \mathbf{P}\Big\{\epsilon_{(n)} - \log(n/2) \ge x\Big\} = 1 - \exp[-\exp(-x)]$$

в упрощенной неасиптотической форме

$$\mathbf{P}\Big\{\epsilon_{(n)} - \log(n/2) \ge x\Big\} \le 2\exp(-x), \tag{1.3.9}$$

которая вытекает неравенства

$$1 - \exp[-\exp(-x)] \le \exp(-x).$$

Это же неравенство следует из выпуклости функции $\exp(x)$ точнее из неравенства $1 - \exp(-z) \le z$.

Неравенство (1.3.9) справедливо при всех x и достаточно больших $n \geq n_\circ$. В данном случае $n_\circ$ не зависит от x. В том, что это утверждение верно, несложно убедиться просматривая доказательство теоремы 1.3.8.

Попробуем теперь выяснить для каких случайных величин справедливо неравенство (1.3.8).

Теорема 1.3.10 *Предположим, что ϵ_i — независимые случайные величины такие, что найдется число $\sigma^2 > 0$, такое, что неравенство*

$$\mathbf{E}\mathrm{e}^{\lambda\epsilon_i} \leq \mathrm{e}^{\sigma^2\lambda^2/2}.$$

выполнено при всех $\lambda > 0$. Тогда при всех $x \geq 0$

$$\mathbf{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i > x\right\} \leq \exp\left\{-\frac{x^2}{2\sigma^2}\right\}.$$

Доказательство. Из неравенства Чернова находим, что

$$\begin{aligned}\mathbf{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i > x\right\} &\leq \min_{\lambda>0}\mathrm{e}^{-\lambda x}\mathbf{E}\exp\left\{\frac{\lambda}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\right\}\\ = \min_{\lambda>0}\mathrm{e}^{-\lambda x}\prod_{i=1}^{n}\mathbf{E}\exp\left\{\frac{\lambda\epsilon_i}{\sqrt{n}}\right\} &\leq \min_{\lambda>0}\mathrm{e}^{-\lambda x}\prod_{i=1}^{n}\exp\left\{\frac{\lambda^2\sigma^2}{n}\right\}\\ = \exp\left\{\min_{\lambda>0}\left[-\lambda x + \frac{\lambda^2\sigma^2}{2}\right]\right\} &= \exp\left\{-\frac{x^2}{2\sigma^2}\right\}. \quad \blacksquare\end{aligned}$$

Случайные величины ϵ_i, для которых неравенство

$$\mathbf{E}\mathrm{e}^{\lambda\epsilon_i} \leq \mathrm{e}^{\lambda^2\sigma^2/2}$$

выполняется при всех $\lambda \in \mathbb{R}$ часто называются *субгауссовскими*. Для этих величин, очевидно, справедлива теорема 1.3.10.

Типичный пример субгауссовских величин — ограниченные случайные величины с нулевым средним.

Теорема 1.3.11 (Хёфдинг) *Пусть ξ — с. в. с нулевым средним $\mathbf{E}\xi = 0$ и $\xi \in [a, b]$. Тогда*

$$\mathbf{E}\exp(\lambda\xi) \leq \exp\left\{\frac{\lambda^2(b-a)^2}{8}\right\}.$$

Доказательство. Рассмотрим функцию

$$\phi(\lambda) = \log\left[\int_a^b e^{\lambda x} p(x)\,dx\right] = \log\left[\mathbf{E}\exp(\lambda\xi)\right]$$

и заметим, что

$$\begin{aligned}
\phi'(\lambda) &= \int_a^b x e^{\lambda x} p(x)\,dx \Big/ \int_a^b e^{\lambda u} p(u)\,du,\\
\phi''(\lambda) &= \int_a^b x^2 e^{\lambda x} p(x)\,dx \Big/ \int_a^b e^{\lambda u} p(u)\,du\\
&\quad - \left\{\int_a^b x e^{\lambda x} p(x)\,dx \Big/ \int_a^b e^{\lambda u} p(u)\,du\right\}^2 = \mathbf{Var}(\zeta),
\end{aligned}$$

где ζ — с. в. с плотностью

$$e^{\lambda x} p(x) \Big/ \int_a^b e^{\lambda u} p(u)\,du.$$

Заметим далее, что $\zeta \in [a,b]$ и что

$$\left[\zeta - \frac{a+b}{2}\right]^2 \le \left(\frac{b-a}{2}\right)^2.$$

Поэтому

$$\mathbf{Var}(\zeta) = \min_x \mathbf{E}\big[\zeta - x\big]^2 \le \left(\frac{b-a}{2}\right)^2.$$

Чтобы завершить доказательство, проинтегрируем неравенство

$$\phi''(\lambda) \le \left(\frac{b-a}{2}\right)^2,$$

учитывая, что $\phi(0) = 0$ и $\phi'(0) = 0$. Тогда находим

$$\phi(\lambda) \le \frac{\lambda^2}{2}\left(\frac{b-a}{2}\right)^2. \quad \blacksquare$$

1.4 Дельта-метод

Ранее мы видели, что для функции распределения $F(x)$ случайной величины ϵ существует и хорошо определен ее эмпирический аналог, а именно, эмпирическая функция распределения выборки

$$\bar{F}(x;\epsilon^n) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\epsilon_i \le x\}. \tag{1.4.10}$$

Для плотности распределения

$$p(x) = \frac{dF(x)}{dx}$$

ее эмпирический аналог уже не является функцией в обычном смысле. Для того чтобы определить эту функцию, довольно естественно продифференцировать $\bar{F}(x;\epsilon^n)$ по x. Тогда из (1.4.10) получим

$$\bar{p}(x;\epsilon^n) = \frac{d\bar{F}(x;\epsilon^n)}{dx} = \frac{1}{n}\sum_{i=1}^{n}\delta(x-\epsilon_i).$$

Здесь $\delta(\cdot)$ — дельта-функция Дирака. Это обобщенная функция, т.е. линейный функционал на пространстве $\mathcal{D}$ бесконечно дифференцируемых функций с финитными носителями. Элементы этого пространства будем обозначать φ, а линейные функционалы от функции f как $\langle f,\varphi\rangle$. Если f — хорошая функция, например, непрерывная, то

$$\langle f,\varphi\rangle = \int_{-\infty}^{\infty} f(x)\varphi(x)\,dx.$$

Дельта-функция Дирака определяется соотношением

$$\langle\delta,\varphi\rangle = \varphi(0).$$

С интуитивной точки зрения δ-функцию можно рассматривать как предел при $n\to\infty$ последовательности

$$\delta_n(x) = \begin{cases} 1/n, & x\in[-1/(2n),1/(2n)] \\ 0, & x\notin[-1/(2n),1/(2n)]. \end{cases}$$

Так как $\delta_n(x)$ интегрируемая функция, то для любой $\varphi\in\mathcal{D}$

$$\lim_{n\to\infty}\langle\delta_n,\varphi\rangle = \lim_{n\to\infty}\frac{1}{n}\int_{-1/(2n)}^{1/(2n)}\varphi(x)\,dx = \varphi(0).$$

В статистике эмпирическая плотность играет исключительно важную роль. Во многих задачах нам нужно оценивать функционалы от плотности распределения. Например,

- *среднее значение* случайной величины

$$M(p) = \int_{-\infty}^{\infty} xp(x)\,dx;$$

- *дисперсия*

$$\sigma^2(p) = \int_{-\infty}^{\infty} [x - M(p)]^2 p(x)\, dx$$
$$= \int_{-\infty}^{\infty} x^2 p(x)\, dx - \left[\int_{-\infty}^{\infty} x p(x)\, dx\right]^2;$$

- характеристическая функция

$$\phi^t(p) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx} p(x)\, dx, \quad \mathrm{i} = \sqrt{-1},$$

- *квантиль* $t^\alpha(p)$ *уровня* α, определяемая как корень уравнения

$$\int_{-\infty}^{t^\alpha(p)} p(x)\, dx = \alpha.$$

Во всех этих примерах мы можем представить интересующий нас объект в виде функционала $\Phi(p)$ от плотности p.

δ-метод это оценка функционала $\Phi(p)$, определяемая следующим образом,

$$\bar{\Phi}(\epsilon^n) = \Phi\left[\bar{p}_{\epsilon^n}\right].$$

Иными словами, что для того, чтобы оценить функционал $\Phi(p)$ надо в него вместо неизвестной плотности p подставить эмпирическую плотность.

Для приведенных выше примеров функционалов мы получим следующие эмпирические версии:

- эмпирическое среднее

$$\bar{M}(\epsilon^n) = \frac{1}{n}\sum_{k=1}^{n} \epsilon_k;$$

- эмпирическая дисперсия

$$\bar{\sigma}^2(\epsilon^n) = \frac{1}{n}\sum_{k=1}^{n} \epsilon_k^2 - \left[\frac{1}{n}\sum_{k=1}^{n} \epsilon_k\right]^2;$$

- эмпирическая характеристическая функция

$$\bar{\phi}^t(\epsilon^n) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{e}^{itX_k};$$

- эмпирический квантиль $\bar{t}^{\alpha}(\epsilon^n)$, который определяется как корень уравнения
$$\#\Big\{k : \epsilon_k \le \bar{t}^{\alpha}(\epsilon^n)\Big\} = \lfloor n\alpha \rfloor,$$
где $\lfloor x \rfloor$ означает целую часть числа x.

Далее мы встретимся с более нетривиальным использованием δ-метода. Его оптимальность обсудим также позднее. В основе доказательства оптимальности δ - метода лежит один из фундаментальных фактов теории вероятностей.

Теорема 1.4.1

$$\lim_{n\to\infty} \sqrt{n}\big[\bar{p}(x;\epsilon^n) - p(x)\big] \stackrel{\mathbf{P}}{=} \sqrt{p(x)}\xi(x) - p(x)\langle \xi, \sqrt{p}\rangle,$$

где $\xi(x)$ — стандартный, белый гауссовский шум.

Стандартный белый гауссовский шум подробно обсуждается в разделе 4.4. Эта теорема ни что иное, как центральная предельная теорема для конечного набора линейных функционалов

$$\bar{\Delta}(\psi_k;\epsilon^n) = \big\langle \psi_k, \sqrt{n}\big(\bar{p}(\cdot;\epsilon^n) - p\big)\big\rangle,\ \psi_k \in \mathcal{D},\ k = 1,\dots,K.$$

Эти функционалы имеют следующий статистический смысл. Пусть нам нужно оценить некоторый линейный функционал от неизвестной плотности распределения p, представимый в виде

$$\Psi = \int_{\mathbb{R}} p(x)\psi(x)\,dx$$

по выборке из независимых, одинаково распределенных случайных величин ϵ^n. Для этого мы используем δ - метод, т.е. оцениваем этот функционал как

$$\bar{\Psi}(\epsilon^n) = \frac{1}{n}\sum_{i=1}^{n} \psi(\epsilon_i)$$

и нас интересует точность этого метода, то есть величина

$$\bar{\Psi}(\epsilon^n) - \Psi = \frac{\bar{\Delta}(\psi;\epsilon^n)}{\sqrt{n}}.$$

Теорема 1.4.1 дает ответ на этот вопрос.

К сожалению, ситуации когда эта теорема может быть использована непосредственно довольно редки. Если говорить очень грубо, то классическая математическая статистика занимается обобщением этой теоремы для

- нелинейных функционалов
- специальных семейств линейных функционалов, число которых K зависит от числа наблюдений.

Отнюдь не все статистические задачи можно решить с помощью обобщений теоремы 1.4.1. Например, никакое ее обобщение невозможно приметить для статистического анализа задачи оценивания параметра μ по независимым наблюдениям

$$Y_i = \mu + \epsilon_i, \quad i = 1, \ldots, n$$

когда ϵ_i равномерно распределены на $[-\sigma, \sigma]$. В тоже время, когда эти величины гауссовские или лапласовские эта теорема прекрасно работает. Это мы увидим позднее.

Глава 2

Расстояния в статистике

Вернемся к нашей статистической модели оценивания скорости света. Мы предполагаем, что данные $Y^n = \{Y_1, \dots, Y_n\}$, полученные в результате измерений распределены как случайные величины

$$Y_i = \mu + \epsilon_i, \quad i = 1, \dots, n,$$

для некоторого неизвестного параметра $\mu \in \mathbb{R}$. Для простоты считаем, что ϵ_i — независимые случайные величины с известной плотностью $p(x)$. Тогда совместная плотность распределения случайных величин $Y^n = \{Y_1, \dots, Y_n\}$ вычисляется как

$$p(y_1, \dots, y_n; \mu) = \prod_{i=1}^{n} p_\epsilon(y_i - \mu), \quad \mu \in \mathbb{R}.$$

На самом деле у данных Y^n, которые имеются в нашем распоряжении, есть своя собственная плотность распределения $p_\circ(x)$, $x \in \mathbb{R}^n$, которую мы, конечно, не знаем. Мы же можем работать только с эмпирической плотностью распределения, которая представляет собой δ-функцию

$$\bar{p}(x; Y^n) = \delta(x - Y^n), \quad x \in \mathbb{R}^n.$$

Таким образом, в пространстве всех плотностей имеются три объекта

- истинная плотность $p_\circ(x)$, которую мы не знаем;
- эмпирическая плотность распределения наблюдений $\delta(x - Y^n)$;
- семейство плотностей $p(x; \mu)$, $\mu \in \mathbb{R}$.

Базовая идея оценивания параметра μ состоит в том, чтобы найти такую плотность из семейства плотностей $p(x;\mu),\ \mu \in \mathbb{R}$, которая как можно ближе была к эмпирической плотности $\delta(x - Y^n)$.

Для того, чтобы решить эту задачу нам по крайней мере нужно определить расстояние или меру близости между плотностями. Если расстояние $d(\cdot,\cdot)$ определено, то идеальная оценка очевидно имеет вид

$$\mu^*(p_\circ) = \arg\min_{\mu} d[p_\circ, p(\cdot;\mu)].$$

Эта величина оценкой, безусловно, не является т.к. зависит от плотности $p_\circ(\cdot)$, которую мы не знаем. Поэтому, нам надо оценить $\mu^*(p_\circ)$ по наблюдениям. Для этого у нас нет никакого выбора кроме δ-метода. То есть мы оцениваем μ следующим образом:

$$\bar{\mu}(Y^n) = \mu^*[\delta(\cdot - Y^n)].$$

2.1 Расстояние полной вариации

Самым естественным расстоянием с вероятностной точки зрения является *расстояние полной вариации*. Для случайных величин ξ, η с плотностями $p_\xi(\cdot)$, $p_\eta(\cdot)$ оно определяется следующим образом:

$$D(p_\xi, p_\eta) = \sup_A \left| \int_A p_\xi(x)\,dx - \int_A p_\eta(x)\,dx \right|;$$

здесь *sup* вычисляется по всем измеримым множествам.

Конечно, на первый взгляд кажется, что использовать это определение крайне неудобно, из-за того, что надо вычислять *sup* вычисляется по всем множествам A. К счастью, имеет место следующий факт:

Теорема 2.1.1 (Шеффе (1947)) *Если существуют плотности распределения $p_\xi(x),\ p_\eta(x)$, $x \in \mathbb{R}^n$ случайных величин $\xi,\ \eta$, то*

$$D(p_\xi, p_\eta) = \frac{1}{2} \int_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)|\,dx.$$

Доказательство. Заметим, что

$$D(p_\xi, p_\eta) = \sup_A \left| \int_A [p_\xi(x) - p_\eta(x)]\,dx \right|.$$

и обозначим для краткости

$$A_\circ = \{x : p_\xi(x) \geq p_\eta(x)\}, \quad A_\circ^c = \{x : p_\xi(x) < p_\eta(x)\}.$$

Тогда

$$D(p_\xi, p_\eta) \geq \int_{A_\circ} [p_\xi(x) - p_\eta(x)]\, dx.$$

С другой стороны,

$$\begin{aligned} \int_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)|\, dx &= \int_{A_\circ} [p_\xi(x) - p_\eta(x)]\, dx \\ - \int_{A_\circ^c} [p_\xi(x) - p_\eta(x)]\, dx &= 2 \int_{A_\circ} [p_\xi(x) - p_\eta(x)]\, dx. \end{aligned} \tag{2.1.1}$$

Следовательно,

$$D(p_\xi, p_\eta) \geq \frac{1}{2} \int_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)|\, dx. \tag{2.1.2}$$

Заметим далее, что для любого измеримого множества A справедливы соотношения

$$\begin{aligned} \left| \int_A [p_\xi(x) - p_\eta(x)]\, dx \right| &= \left| \int_{A \cap A_\circ} [p_\xi(x) - p_\eta(x)]\, dx \right. \\ &\qquad \left. + \int_{A \cap A_\circ^c} [p_\xi(x) - p_\eta(x)]\, dx \right| \\ &= \left| \int_{A \cap A_\circ} [p_\xi(x) - p_\eta(x)]\, dx - \int_{A \cap A_\circ^c} [p_\eta(x) - p_\xi(x)]\, dx \right| \\ &\leq \max\left\{ \int_{A \cap A_\circ} [p_\xi(x) - p_\eta(x)]\, dx, \int_{A \cap A_\circ^c} [p_\eta(x) - p_\xi(x)]\, dx \right\} \\ &\leq \max\left\{ \int_{A_\circ} [p_\xi(x) - p_\eta(x)]\, dx, \int_{A_\circ^c} [p_\eta(x) - p_\xi(x)]\, dx \right\} \\ &= \max\left\{ \int_{A_\circ} [p_\xi(x) - p_\eta(x)]\, dx, \int_{A_\circ} [p_\xi(x) - p_\eta(x)]\, dx \right\}. \end{aligned}$$

Таким образом, мы видим что верхняя граница для расстояния полной вариации совпадает с нижней границей (2.1.2). ■

2.2 Расстояние Кульбака-Лейблера

Расстояние полной вариации очень естественная и хорошая мера близости между плотностью $p_\circ(\cdot)$ и плотностями $p(\cdot;\mu)$. Однако использовать его достаточно сложно поскольку непонятно как вычислить

$$\mu^*(p_\circ) = \arg\min_\mu D[p_\circ, p(\cdot;\mu)]$$

Кроме того ясно, что

$$D[\delta(\cdot - Y^n), p(\cdot;\mu)] = \frac{1}{2}\int_{\mathbb{R}^n} |\delta(x - Y^n) - p(x;\mu)|\, dx = 1$$

для любой ограниченной плотности $p(\cdot;\mu)$.

Реализуемый подход к решению задачи аппроксимации неизвестного распределения распределениями из заданного семейства основан на другом расстоянии, а именно, на *расстоянии Кульбака-Лейблера*

$$K(p_\xi, p_\eta) = \int_{\mathbb{R}^n} p_\xi(x) \log \frac{p_\xi(x)}{p_\eta(x)}\, dx.$$

Первое, что бросается в глаза это то, что $K(p_\xi, p_\eta)$ не является расстоянием в обычном смысле поскольку $K(p_\xi, p_\eta) \neq K(p_\eta, p_\xi)$. Хотя с наивной точки зрения отсутствие симметрии кажется абсурдным фактом, но если немного подумать, то становится ясно, что это реальная ситуация в статистике. Действительно, с одной стороны у есть реальные данные, плотность распределения которых нам неизвестна и которой мы никаким образом не можем управлять. В терминах расстояния Кульбака-Лейблера это значит, что мы берем $p_\xi(x) = p_\circ(x)$. С другой стороны, у нас имеется семейство распределений $p(x;\mu)$, $\mu \in \mathbb{R}$. В этом семействе мы может выбирать любую плотность. Понятно, что данная ситуация не является симметричной.

Если решено пользоваться расстоянием Кульбака-Лейблера для подгонки вероятностной модели к реальным данным, то прежде всего нам нужно понять как связано это расстояние с расстоянием полной вариации, которое является безусловно самой естественной мерой близости искусственной модели и реальных данных.

Теорема 2.2.1 (Неравенство Пинскера) *Если случайные величины ξ, η имеют плотности $p_\xi(x)$, $p_\eta(x)$, $x \in \mathbb{R}^n$, то*

$$K(p_\xi, p_\eta) \geq 2[D(p_\xi, p_\eta)]^2.$$

Доказательство. Как и в доказательстве теоремы Шеффе обозначим

$$A_\circ = \{x : p_\xi(x) \geq p_\eta(x)\}, \quad A_\circ^c = \{x : p_\xi(x) < p_\eta(x)\}$$

и заметим, что в силу теоремы Шеффе 2.1.1

$$\begin{aligned} D(p_\xi, p_\eta) &= \frac{1}{2}\int_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)|\,dx \\ &= \frac{1}{2}\int_{A_\circ} [p_\xi(x) - p_\eta(x)]\,dx + \frac{1}{2}\int_{A_\circ^c} [p_\eta(x) - p_\xi(x)]\,dx \\ &= \left[\int_{A_\circ} p_\xi(x)\,dx - \int_{A_\circ} p_\eta(x)\,dx\right] = [P_\xi(A_\circ) - P_\eta(A_\circ)]; \end{aligned} \tag{2.2.3}$$

здесь

$$P_\xi(A_\circ) = \int_{A_\circ} p_\xi(x)\,dx, \quad P_\eta(A_\circ) = \int_{A_\circ} p_\eta(x)\,dx.$$

Для того чтобы получить нижнюю границу для расстояния Кульбака-Лейблера воспользуемся выпуклостью функции $-\log(x)$.

Точнее, мы будем использовать *неравенство Йенсена*

$$\int p(x) f[q(x)]\,dx \geq f\left[\int p(x) q(x)\,dx\right],$$

где $p(x)$ — вероятностная плотность распределения, а $f(\cdot)$ — выпуклая функция.

В нашем случае, взяв $f(y) = -\log(y)$ и применив неравенство Йенсена получим

$$\begin{aligned} K(p_\xi, p_\eta) &= \int_{A_\circ} p_\xi(x) \log\frac{p_\xi(x)}{p_\eta(x)}\,dx + \int_{A_\circ^c} p_\xi(x) \log\frac{p_\xi(x)}{p_\eta(x)}\,dx \\ &= -P_\xi(A_\circ)\int_{A_\circ} \frac{p_\xi(x)}{P_\xi(A_\circ)} \log\frac{p_\eta(x)}{p_\xi(x)}\,dx \\ &\quad -P_\xi(A_\circ^c)\int_{A_\circ^c} \frac{p_\xi(x)}{P_\xi(A_\circ^c)} \log\frac{p_\eta(x)}{p_\xi(x)}\,dx \\ &\geq -P_\xi(A_\circ)\log\left[\frac{1}{P_\xi(A_\circ)}\int_{A_\circ} p_\eta(x)\,dx\right] \\ &\quad -P_\xi(A_\circ^c)\log\left[\frac{1}{P_\xi(A_\circ^c)}\int_{A_\circ^c} p_\eta(x)\,dx\right] \\ &= -P_\xi(A_\circ)\log\left[\frac{P_\eta(A_\circ)}{P_\xi(A_\circ)}\right] - P_\xi(A_\circ^c)\log\left[\frac{P_\eta(A_\circ^c)}{P_\xi(A_\circ^c)}\right] \\ &-P_\xi(A_\circ)\log\left[\frac{P_\eta(A_\circ)}{P_\xi(A_\circ)}\right] - [1 - P_\xi(A_\circ)]\log\left[\frac{1 - P_\eta(A_\circ)}{1 - P_\xi(A_\circ)}\right]. \end{aligned} \tag{2.2.4}$$

Обозначим для краткости

$$p = P_\xi(A_\circ), \quad q = P_\eta(A_\circ).$$

В этих обозначениях тождество (2.2.3) и неравенство (2.2.4) запишутся следующим образом:

$$\begin{aligned} &D(p_\xi, p_\eta) = p - q \\ &K(p_\xi, p_\eta) \geq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}. \end{aligned} \tag{2.2.5}$$

Рассмотрим следующую функцию

$$f(q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - \lambda (p-q)^2$$

и заметим, что ее производная равна

$$f'(p) = -\frac{p}{q} + \frac{1-p}{1-q} + 2\lambda(p-q) = (q-p)\left[\frac{1}{q(1-q)} - 2\lambda\right].$$

Поскольку для любого $q \in [0,1]$

$$\frac{1}{q(1-q)} \geq 4,$$

то при любом $\lambda < 2$ производная $f'(q)$ обращается в нуль только в точке $q = p$, причем $f(p) = 0$. Заметим еще, что $f(0) = +\infty$ и $f(1) = +\infty$. Это означает, что функция $f(q)$ достигает минимума в точке $q = p$, то есть

$$f(q) \geq f(p) = 0.$$

Поэтому, пользуясь определением функции $f(q)$ получаем, что

$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq 2(p-q)^2.$$

Отсюда и из (2.2.5) получаем.

$$K(p_\xi, p_\eta) \geq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq 2(p-q)^2 = 2[D(p_\xi, p_\eta)]^2. \quad \blacksquare$$

У расстояния Кульбака-Лейблера есть замечательное свойство, которое делает его очень удобным для использования.

Теорема 2.2.2 *Пусть $\xi^n = \{\xi_1, \ldots, \xi_n\}$ и $\eta^n = \{\eta_1, \ldots, \eta_n\}$ случайные векторы образованные из независимых, одинаково распределенных случайных величин. Тогда*

$$K(p_{\xi^n}, p_{\eta^n}) = nK(p_{\xi_1}, p_{\eta_1}).$$

Доказательство. Пользуясь независимостью случайных величин, находим

$$\begin{aligned}
K(p_{\xi^n}, p_{\eta^n}) &= \int_{\mathbb{R}^n} p_{\xi^n}(y_1, \ldots, y_n) \log \frac{p_{\xi^n}(y_1, \ldots, y_n)}{p_{\eta^n}(y_1, \ldots, y_n)}\, dy_1 \cdots y_n \\
&= \int_{\mathbb{R}^n} p_{\xi^n}(y_1, \ldots, y_n) \sum_{i=1}^{n} \log \frac{p_{\xi_1}(y_i)}{p_{\eta_1}(y_i)}\, dy_1 \cdots dy_n \\
&= \sum_{i=1}^{n} \int_{\mathbb{R}^n} p_{\xi^n}(y_1, \ldots, y_n) \log \frac{p_{\xi_1}(y_i)}{p_{\xi_1}(y_i)}\, dy_1 \cdots dy_n \\
&= n \int_{\mathbb{R}^1} p_{\xi_1}(x) \log \frac{p_{\xi_1}(x)}{p_{\eta_1}(x)}\, dx = nK(p_{\xi_1}, p_{\eta_1}). \quad \blacksquare
\end{aligned}$$

В статистической литературе можно встретить различные расстояния, например, такие как расстояние Хеллингера и расстояние χ-квадрат. Эти расстояния тесно связаны с расстояниями по вариации и расстоянием Кульбака-Лейблера.

Для того, чтобы определить расстояние Хеллингера, заметим, что из неравенства Коши-Буняковского сразу же вытекает следующая верхняя граница для расстояния полной вариации:

$$\begin{aligned}
D(p_\xi, p_\eta) =& \frac{1}{2} \int_{\mathbb{R}^n} |p_\xi(x) - p_\eta(x)|\, dx \\
=& \frac{1}{2} \int \left(\sqrt{p_\xi(x)} + \sqrt{p_\eta(x)} \right) \left| \sqrt{p_\xi(x)} - \sqrt{p_\eta(x)} \right| dx \\
\le& \frac{1}{2} \left[\int_{\mathbb{R}^n} \left(\sqrt{p_\xi(x)} + \sqrt{p_\eta(x)} \right)^2 dx \right]^{1/2} \\
& \times \left[\int_{\mathbb{R}^n} \left| \sqrt{p_\xi(x)} - \sqrt{p_\eta(x)} \right|^2 dx \right]^{1/2} \\
\le& \left[\int_{\mathbb{R}^n} \left| \sqrt{p_\xi(x)} - \sqrt{p_\eta(x)} \right|^2 dx \right]^{1/2}.
\end{aligned}$$

Правую часть в этом неравенстве, а именно, величину

$$H(p_\xi, p_\eta) \stackrel{\text{def}}{=} \left[\int_{\mathbb{R}^n} \left| \sqrt{p_\xi(x)} - \sqrt{p_\eta(x)} \right|^2 dx \right]^{1/2}$$

называют расстоянием Хеллингера.

Расстояние χ-квадрат является первым нетривиальным членом разложения Тейлора для расстояния Кульбака-Лейблера при близких плотностях $p_\xi(x) \approx p_\eta(x)$. В этом случае, пользуясь тем, что $\log(1 + x) \approx$

$x - x^2/2$ при малых $|x|$, находим

$$\begin{aligned} K(p_\xi, p_\eta) &= -\int_{\mathbb{R}^n} p_\xi(x) \log\left[1 + \frac{p_\eta(x)}{p_\xi(x)} - 1\right] dx \\ &\approx -\int_{\mathbb{R}^n} p_\xi(x)\left[\frac{p_\eta(x)}{p_\xi(x)} - 1\right] dx + \frac{1}{2}\int p_\xi(x)\left[\frac{p_\eta(x)}{p_\xi(x)} - 1\right]^2 dx \\ &= \frac{1}{2}\int_{\mathbb{R}^n} \frac{[p_\xi(x) - p_\eta(x)]^2}{p_\xi(x)}\, dx. \end{aligned}$$

Это соотношение лежит в основе определения расстояния χ-квадрат

$$\chi^2(p_\xi, p_\eta) \stackrel{\text{def}}{=} \int_{\mathbb{R}^n} \frac{[p_\xi(x) - p_\eta(x)]^2}{p_\xi(x)}\, dx.$$

Конечно, можно придумать много других расстояний между плотностями распределений, но в действительности, самым полезным и самым используемыми расстоянием в статистике является расстояние Кульбака-Лейблера. Почему это так мы увидим далее.

Глава 3

Метод максимального правдоподобия

Принцип максимума правдоподобия является одним из основополагающих методов статистики. Практически все разумные статистические методы и алгоритмы основаны на нем или на его модификациях.

3.1 Принцип максимального правдоподобия

Предположим, что мы решили использовать для подгонки вероятностной модели расстояние Кульбака-Лейблера.

Если бы была известна точно плотность распределения наблюдений, то мы бы выбрали наилучшую плотность в семействе или, что эквивалентно, параметр μ следующим образом:

$$\mu^*(p_\circ) = \arg\min_{\mu} \int_{\mathbb{R}^n} p_\circ(z) \log \frac{p_\circ(z)}{p(z;\mu)}\, dz.$$

Легко видеть, что

$$\mu^*(p_\circ) = \arg\max_{\mu} \int_{\mathbb{R}^n} p_\circ(z) \log[p(z;\mu)]\, dz.$$

Заметим тем не менее, что для того чтобы пользоваться этой формулой нужно знать плотность $p_\circ(z)$. Однако, заметим также, что

$$L[p_\circ, \mu] = \int_{\mathbb{R}^n} p_\circ(z) \log[p(z;\mu)]\, dz$$

представляет собой линейный функционал от плотности $p_\circ(\cdot)$.

Основная идея, на которой базируется метод максимального правдоподобия, состоит в том, чтобы оценить этот функционал на основе имеющихся данных Y^n с помощью δ-метода. А именно, использовать следующую оценку

$$\bar{L}(\mu; Y^n) = \int_{\mathbb{R}^n} \delta(z - Y^n) \log\bigl[p(z;\mu)\bigr]\, dz = \log\bigl[p(Y^n;\mu)\bigr].$$

Эта функция называется *логарифмом правдоподобия*, а

$$\bar{\mu}(Y^n) = \arg\max_{\mu}\bigl\{\log[p(Y^n;\mu)]\bigr\}$$

— *оценкой максимального правдоподобия* параметра μ.

Метод максимального правдоподобия предложил Р. Фишер когда ему было 22 года.

Для статистической модели оценивания параметра сдвига по независимым наблюдениям

$$Y_i = \mu + \epsilon_i, \quad i = 1, \dots, n$$

очевидно, имеем

$$\bar{L}(\mu; Y^n) = \sum_{i=1}^{n} \log[p(Y_i - \mu)].$$

Поэтому оценка максимального правдоподобия вычисляется как

$$\bar{\mu}(Y^n) = \arg\max_{\mu}\left\{\sum_{i=1}^{n} \log[p_\epsilon(Y_i - \mu)]\right\}.$$

Заметим, что если случайные величины $Y_i,\ i = 1, \dots, n$ независимы, одинаково распределены и

$$\mathbf{E}\bigl|\log[p(Y_1 - \tilde{\mu})]\bigr| < \infty,$$

для любого $\tilde{\mu}$, то в силу закона больших чисел при фиксированном $\tilde{\mu}$ и $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} \log[p(Y_i - \tilde{\mu})] \to \mathbf{E}\log\bigl[p(Y_1 - \tilde{\mu})\bigr] = \int_{\mathbb{R}^1} p_\circ(x) \log[p(x - \tilde{\mu})]\, dx,$$

где $p_\circ(x)$ — истинная плотность распределения случайной величины Y_1. Поэтому мы можем ожидать, что

$$\begin{aligned}\bar{\mu}(Y^n) \to \arg\max_{\tilde{\mu}} \mathbf{E}\log\bigl[p(Y_1 - \tilde{\mu})\bigr] &= \arg\min_{\tilde{\mu}} \mathbf{E}\log\frac{p_\circ(Y_1)}{p(Y_1 - \tilde{\mu})} \\ &= \arg\min_{\tilde{\mu}} K[p_\circ, p(\cdot - \tilde{\mu})].\end{aligned}$$

Таким образом, при больших n метод максимального правдоподобия выбирает плотность из заданного параметрического семейства, находящуюся на минимальном расстоянии Кульбака-Лейблера от истинной, но неизвестной плотности распределения наблюдений.

Далее мы подробно рассмотрим вопрос об оптимальности этого метода. Сейчас же посмотрим как вычисляются оценки максимального правдоподобия в простейших случаях.

3.2 Оценивание параметра сдвига

3.2.1 Гауссовские ошибки

Предположим, что ошибки ϵ_i в статистической модели

$$Y_i = \mu + \epsilon_i, \quad i = 1, \dots, n$$

являются гауссовскими с нулевым средним, точнее имеют плотность

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Тогда логарифм отношения правдоподобия вычисляется как

$$\bar{L}(\tilde{\mu}; Y^n) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i - \tilde{\mu}]^2 - \frac{n \log(2\pi\sigma^2)}{2}.$$

Для того, чтобы найти оценку максимального правдоподобия нам нужно найти величину $\bar{\mu}(Y^n)$, которая максимизирует квадратичную функцию $\bar{L}_{Y^n}(\tilde{\mu})$. Очевидно, что $\bar{\mu}(Y^n)$ является корнем уравнения

$$\left.\frac{d\bar{L}(\mu; Y^n)}{d\mu}\right|_{\mu=\bar{l}_{X^n}(\tilde{\mu})} = \sum_{i=1}^{n} [Y_i - \bar{\mu}(Y^n)] = 0.$$

Отсюда сразу же найдем

$$\bar{\mu}(Y^n) = \bar{Y}^n = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

То есть оценкой максимального правдоподобия параметра μ при гауссовских ошибках является эмпирическое среднее.

Вероятностный анализ эмпирического среднего трудностей не представляет. При его проведении мы можем отказаться от гипотезы о гауссовости ошибок, заменив ее на условие

$$\mathbf{E}\epsilon_i = 0, \quad \mathbf{E}\epsilon_i^2 = \sigma^2.$$

При этом мы сохраним для простоты выкладок предположение о независимости случайных величин ϵ_i. Тогда эмпирическое среднее имеет следующие свойства:

- *Среднее оценки*
$$\mathbf{E}\bar{Y}^n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(\mu + \epsilon_i) = \mu.$$

- *Среднеквадратичный риск*
$$\mathbf{E}(\bar{Y}^n - \mu)^2 = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right]^2 = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2 = \frac{\sigma^2}{n}. \quad (3.2.1)$$

- *Предельное распределение ошибки оценивания.* Если $\mathbf{E}|\epsilon_i|^3 < \infty$, то
$$\lim_{n\to\infty}\mathbf{P}\left\{\frac{\sqrt{n}|\bar{Y}^n - \mu|}{\sigma} \geq z\right\} = \frac{2}{\sqrt{2\pi}}\int_z^\infty \mathrm{e}^{-u^2/2}\,du.$$

  Заметим, что если ошибки гауссовские, то предел в этой формуле можно опустить.

Таким образом, как и в случае эмпирической функции распределения мы видим, что оценка приближается к истинному значению параметра со скоростью $\sigma/\sqrt{n}$.

В статистике наряду с оценкой параметра, как правило, требуется контроль риска оценки. По этим как правило понимается оценивание величины

$$\mathbf{E}(\bar{Y}^n - \mu)^2.$$

Иными словами, мы хотим, если это возможно, знать не только оценку параметра, но и точность нашего метода оценивания.

В рассматриваемом случае риск полностью определяется величиной σ^2, которая есть не что иное, как дисперсия наблюдений

$$\sigma^2 = \mathbf{E}(Y_i - \mathbf{E}Y_i)^2 = \mathbf{E}Y_i^2 - (\mathbf{E}Y_i)^2.$$

Для ее оценивания будем использовать δ-метод, который дает следующую оценку дисперсии

$$\bar{\sigma}^2(Y^n) = \frac{1}{n}\sum_{i=1}^n Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^n Y_i\right)^2.$$

Нетрудно проверить, что эту оценку можно записать в следующей эквивалентной форме

$$\bar{\sigma}^2(Y^n) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \frac{1}{n}\sum_{s=1}^n Y_s\right)^2.$$

Также нетрудно понять, что это очень хорошая оценка параметра σ. Действительно, подставив в нее $Y_i = \mu + \epsilon_i$, найдем

$$\begin{aligned}
\bar{\sigma}^2(Y^n) &= \frac{1}{n}\sum_{i=1}^n \left(\epsilon_i - \frac{1}{n}\sum_{k=1}^n \epsilon_k\right)^2 = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \left(\frac{1}{n}\sum_{k=1}^n \epsilon_k\right)^2 \\
&= \sigma^2 + \frac{1}{n}\sum_{i=1}^n [\epsilon_i^2 - \sigma^2] - \left(\frac{1}{n}\sum_{k=1}^n \epsilon_k\right)^2 \\
&= \sigma^2 + \sigma^2\left\{\frac{1}{n}\sum_{i=1}^n \left(\frac{\epsilon_i^2}{\sigma^2} - 1\right) - \left(\frac{1}{n}\sum_{k=1}^n \frac{\epsilon_k}{\sigma}\right)^2\right\} \\
&= \sigma^2 + \frac{\sigma^2}{\sqrt{n}}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{\epsilon_i^2}{\sigma^2} - 1\right) - \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_{k=1}^n \frac{\epsilon_k}{\sigma}\right)^2\right\}.
\end{aligned} \tag{3.2.2}$$

Заметим, что

$$\mathbf{E}\bar{\sigma}^2(Y^n) = \sigma^2 - \frac{1}{n^2}\mathbf{E}\sum_{i,k=1}^n \epsilon_i\epsilon_k = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\left(1 - \frac{1}{n}\right).$$

Поэтому, часто вместо оценки $\bar{\sigma}^2(Y^n)$ используют ее несмещенную версию

$$\hat{\sigma}^2(Y^n) = \bar{\sigma}^2(Y^n)\left(1 - \frac{1}{n}\right)^{-1} = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}^n)^2.$$

Асимптотические свойства оценки дисперсии описывает следующая теорема.

Теорема 3.2.1 *Если* $\mathbf{E}\epsilon_i^6 < \infty$, *то*

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{\sqrt{n}(\bar{\sigma}^2(Y^n) - \sigma^2)}{\sigma^2\rho} \le x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x \mathrm{e}^{-u^2/2}\,du, \tag{3.2.3}$$

где

$$\rho^2 = \mathbf{E}\left[\left(\frac{\epsilon_i}{\sigma}\right)^2 - 1\right]^2.$$

Доказательство. В силу центральной предельной теоремы

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{1}{\rho\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\epsilon_i^2}{\sigma^2} - 1\right) \le x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \mathrm{e}^{-u^2/2}\,du,$$

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{\epsilon_k}{\sigma} \le x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \mathrm{e}^{-u^2/2}\,du.$$

Из этих соотношений и (3.2.2) непосредственно вытекает (3.2.3). ■

Определение 3 *Случайная величина η имеет распределение χ-квадрат с k степенями свободы если она представима в виде*

$$\eta = \sum_{i=1}^{k} \xi_i^2,$$

где ξ_i — независимые, стандартные гауссоавские случайные величины.

Завершает этот раздел следующий любопытный факт.

Теорема 3.2.2 *Если $Y_i,\ i = 1\ldots, n$ — независимые, одинаково распределенные гауссовские случайные величины, то оценки среднего $\bar{Y}^n$ и дисперсии $\bar{\sigma}^2(Y^n)$ — независимые случайные величины. Кроме того, случайная величина $n\bar{\sigma}^2(Y^n)/\sigma^2$ имеет распределение χ-квадрат с $n-1$ степенью свободы.*

Доказательство. Оно основано на том, что если гауссовские случайные величины ξ и η_k, $k = 1, \ldots, n$ некорреликованны, т.е.

$$\mathbf{E}\big[\xi - \mathbf{E}\xi\big]\cdot\big[\eta_k - \mathbf{E}\eta_k\big] = 0, \quad k = 1, \ldots, n,$$

то они независимы, т.е. для любой измеримой функции $f : \mathbb{R}^n \to \mathbb{R}$

$$\mathbf{P}\Big\{\xi \le x; f(\eta_1, \ldots, \eta_n) \le y\Big\} = \mathbf{P}\Big\{\xi \le x\Big\}\mathbf{P}\Big\{f(\eta_1, \ldots, \eta_n) \le y\Big\}.$$

Возьмем

$$\xi = \bar{Y}^n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \mu + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i,$$

$$\eta_k = Y_k - \bar{Y}^n = \epsilon_k - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i.$$

Заметив, что

$$\mathbf{E}[\xi - \mathbf{E}\xi]\eta_k = \mathbf{E}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left[\epsilon_k - \frac{1}{n}\sum_{k=1}^{n}\epsilon_k\right] = \mathbf{E}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\epsilon_k - \mathbf{E}\frac{1}{n^2}\sum_{i,k=1}^{n}\epsilon_i\epsilon_k$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\epsilon_i\epsilon_k - \frac{1}{n^2}\sum_{i,k=1}^{n}\mathbf{E}\epsilon_i\epsilon_k = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$$

Для доказательства того, что случайная величина $n\bar{\sigma}^2(Y^n)/\sigma^2$ имеет распределение χ-квадрат с $n-1$ степенью свободы заметим, что

$$\frac{n\bar{\sigma}^2(Y^n)}{\sigma^2} = \left\|\frac{\epsilon^n}{\sigma} - \left\langle\frac{\epsilon^n}{\sigma}, e_1\right\rangle e_1\right\|^2, \tag{3.2.4}$$

где $\|\cdot\|$ и $\langle\cdot,\cdot\rangle$ — евклидова норма и скалярное произведение в $\mathbb{R}^n$, а e_1 — вектор с компонентами

$$e_1 = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)^\top$$

Пусть $e_2, \ldots, e_n$ — некоторая система ортонормальных векторов, ортогональных e_1. Тогда

$$\frac{\epsilon^n}{\sigma} = \sum_{i=1}^{n}\left\langle\frac{\epsilon^n}{\sigma}, e_i\right\rangle e_i.$$

Поэтому из (3.2.4) находим

$$\frac{n\bar{\sigma}^2(Y^n)}{\sigma^2} = \sum_{i=2}^{n}\left\langle\frac{\epsilon^n}{\sigma}, e_i\right\rangle^2.$$

Для завершения доказательства теоремы осталось заметить, что $\langle\epsilon^n/\sigma, e_i\rangle$ — независимые, стандартные гауссовские случайные величины. ■

Как мы уже видели метод максимального правдоподобия минимизирует расстояние Кульбака-Лейблера между соответствующими мерами. Посмотрим, что означает это свойство для модели параметра сдвига при гауссовских ошибках. Обозначим для краткости

$$p^n(x_1, \ldots, x_n) = \prod_{i=1}^{n}p(x_i), \text{ где } p(x) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{x^2}{2}\right].$$

Воспользовавшись теоремой 2.2.2, имеем

$$\begin{aligned} &K[p^n(\cdot-\mu_1),p^n(\cdot-\mu_2)] = nK[p(\cdot-\mu_1),p(\cdot-\mu_2)] \\ &= \frac{n}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} \mathrm{e}^{-(x-\mu_1)^2/(2\sigma^2)}\left[-\frac{(x-\mu_1)^2}{2\sigma^2}+\frac{(x-\mu_2)^2}{2\sigma^2}\right]dx \\ &= \frac{n}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{-u^2/2}\left[-\frac{u^2}{2}+\frac{(\sigma u+\mu_1-\mu_2)^2}{2\sigma^2}\right]dx \\ &= \frac{n}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{-u^2/2}\frac{(\mu_1-\mu_2)^2}{2\sigma^2}\,dx = \frac{n(\mu_1-\mu_2)^2}{2\sigma^2}. \end{aligned} \tag{3.2.5}$$

То есть расстояние Кульбака-Лейблера пропорционально $(\mu_1-\mu_2)^2$. Любопытное и очень полезное свойство оценки максимального правдоподобия $\bar{\mu}$ вытекает из (3.2.5) и (3.2.1) и имеет вид следующего тождества:

$$\mathbf{E}K[p^n(\cdot-\bar{\mu}),p^n(\cdot-\mu)] = \frac{1}{2},$$

которое справедливо в случае когда ошибки гауссовские. То есть среднее значение расстояния Кульбака-Лейблера между истинной плотностью распределения и ее оценкой $p^{\bar{\mu}}_{Y^n}(x)$, $x\in\mathbb{R}^n$ не зависит ни от числа наблюдений n, ни уровня шума σ. Также понятно, что

$$\mathbf{P}\Big\{K[p^n(\cdot-\bar{\mu}),p^n(\cdot-\mu)]\geq z\Big\} = \mathbf{P}\left\{\frac{\xi^2}{2}\geq z\right\} = \frac{2}{\sqrt{2\pi}}\int_{\sqrt{2z}}^{\infty}\mathrm{e}^{-u^2/2}\,du.$$

3.2.2 Лапласовские ошибки

Теперь рассмотрим что будет происходить если в модели оценивания параметра сдвига

$$Y_i = \mu+\epsilon_i,\quad i=1,\dots,n$$

ошибки, которые являются независимыми случайными величинами имеют распределение Лапласа с плотностью

$$p(x) = \frac{1}{2\sigma}\mathrm{e}^{-|x|/\sigma},\quad \sigma>0.$$

В этом случае логарифм отношения правдоподобия имеет следующий вид:

$$\bar{L}(\tilde{\mu};Y^n) = -\frac{1}{\sigma}\sum_{i=1}^{n}|Y_i-\tilde{\mu}| - n\log(2\sigma)$$

а оценка максимального правдоподобия вычисляется как

$$\bar{\mu}(Y^n) = \arg\min_{\tilde{\mu}} \sum_{i=1}^{n} |Y_i - \tilde{\mu}|.$$

В зависимости от того является число n четным или нечетным $\bar{\mu}(Y^n)$ может быть единственным или неединственным числом. Если $n = 2\lfloor n/2 \rfloor + 1$ - нечетное число, то легко проверить что корень уравнения

$$\frac{d}{d\tilde{\mu}} \sum_{i=1}^{n} |Y_i - \tilde{\mu}| = \sum_{i=1}^{n} \operatorname{sign}(Y_i - \tilde{\mu}) = 0$$

будет равен

$$\bar{\mu}(Y^n) = Y_{((n-1)/2+1)},$$

а если n - четное, то в качестве корня можно взять любое

$$\bar{\mu}(Y^n) \in [Y_{(n/2)}, Y_{(n/2+1)}].$$

То есть, оценка $\bar{\mu}(Y^n)$ является эмпирической медианой выборки Y^n. Для простоты далее будем считать, что

$$\bar{\mu}(Y^n) = Y_{(\lfloor n/2 \rfloor)}.$$

Для того, чтобы изучить свойства медианы, применим тот же метод, который был использован при исследовании экстремальных значений выборок. Точнее, воспользуемся методом инверсии

$$\bar{\mu}_{Y^n} = F_Y^{-1}(U_{(\lfloor n/2 \rfloor)})$$

и теоремой Пайка

$$U_{(\lfloor n/2 \rfloor)} \stackrel{\mathbf{P}}{=} \sum_{i=1}^{\lfloor n/2 \rfloor} \zeta_i \Big/ \sum_{k=1}^{n+1} \zeta_k.$$

Из формулы Тейлора и центральной предельной теоремы найдем

$$
\begin{aligned}
U_{(\lfloor n/2\rfloor)} &= \left[\frac{1}{2}+\frac{1}{n}\sum_{i=1}^{n/2}(\zeta_i-1)\right]\Big/\left[1+\frac{1}{n}\sum_{k=1}^{n+1}(\zeta_k-1)\right]\\
&= \left[\frac{1}{2}+\frac{1}{n}\sum_{i=1}^{n/2}(\zeta_i-1)\right]\cdot\left[1-\frac{1}{n}\sum_{k=1}^{n+1}(\zeta_k-1)+O\left(\frac{1}{n}\right)\right]\\
&= \frac{1}{2}+\frac{1}{n}\sum_{i=1}^{n/2}(\zeta_k-1)-\frac{1}{2n}\sum_{k=1}^{n+1}(\zeta_k-1)+O\left(\frac{1}{n}\right) \qquad (3.2.6)\\
&= \frac{1}{2}+\frac{1}{2n}\sum_{i=1}^{n/2}(\zeta_k-1)-\frac{1}{2n}\sum_{k=n/2+1}^{n+1}(\zeta_k-1)+O\left(\frac{1}{n}\right)\\
&= \frac{1}{2}+\frac{1}{2\sqrt{n}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n+1}\xi_k\right]+O\left(\frac{1}{n}\right),
\end{aligned}
$$

где

$$
\xi_k=\begin{cases}\zeta_k-1, & k\le n/2,\\ 1-\zeta_k, & n/2<k\le n+1.\end{cases}
$$

Далее заметив, что в поскольку $\mathbf{E}\xi_k^2=1$, то согласно центральной предельной теореме

$$
\lim_{n\to\infty}\mathbf{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n+1}\xi_k\le x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-u^2/2}\,du
$$

и, следовательно,

$$
\lim_{n\to\infty}\mathbf{P}\left\{2\sqrt{n}\left(U_{(\lfloor n/2\rfloor)}-\frac{1}{2}\right)\le x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-u^2/2}\,du. \qquad (3.2.7)
$$

Теперь используем этот факт и формулу Тейлора. При этом будем предполагать, что плотность $p_Y(x)$ распределения величины Y имеет при всех x ограниченную производную. Тогда находим

$$
\begin{aligned}
F_Y^{-1}(U_{(\lfloor n/2\rfloor)}) &= F_Y^{-1}\left(\frac{1}{2}+U_{(\lfloor n/2\rfloor)}-\frac{1}{2}\right)\\
&= F_Y^{-1}\left(\frac{1}{2}\right)+\frac{1}{p_Y[F_Y^{-1}(1/2)]}\left(U_{(\lfloor n/2\rfloor)}-\frac{1}{2}\right)\\
&\qquad +O\left[\left(U_{(\lfloor n/2\rfloor)}-\frac{1}{2}\right)^2\right] \qquad (3.2.8)\\
&= F_Y^{-1}\left(\frac{1}{2}\right)+\frac{1}{p_Y[F_Y^{-1}(1/2)]}\left(U_{(\lfloor n/2\rfloor)}-\frac{1}{2}\right)+O\left(\frac{1}{n}\right).
\end{aligned}
$$

Предположим, что плотность распределения ошибок симметрична

$$p(x) = p(-x), \quad x \in \mathbb{R}.$$

Тогда ясно, что

$$F(0) = \frac{1}{2} \text{ и поскольку } F_Y(z) = F(x-\mu),$$

то

$$F_Y(\mu) = \frac{1}{2}, \quad F_Y^{-1}\left(\frac{1}{2}\right) = \mu, \quad p_Y[F_Y^{-1}(1/2)] = p_\epsilon(0).$$

Отсюда и из (3.2.7), (3.2.8) приходим к следующему результату

Теорема 3.2.3 *Если плотность $p(\cdot)$ симметрична и имеет во всех точках ограниченную производную, то*

$$\lim_{n\to\infty} \mathbf{P}\Big\{2p_\epsilon(0)\sqrt{n}\big[\bar{\mu}(Y^n) - \mu\big] \le x\Big\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-u^2/2}\, du$$

и

$$\lim_{n\to\infty} n\mathbf{E}\big[\bar{\mu}(Y^n) - \mu\big]^2 = \frac{1}{4p_\epsilon^2(0)}.$$

Эта теорема проливает свет на так называемую робастность медианы. Под робастностью понимается устойчивость оценок по отношению к выбору модели. К сожалению, мы никогда не знаем точно какова на самом деле плотность распределения ошибок. Поэтому посмотрим как поведут себя среднее и медиана в ситуации когда ошибки распределены по закону Коши с плотностью

$$p(x) = \frac{1}{\pi\sigma[1 + (x/\sigma)^2]}.$$

Хорошо известно и легко проверить, вычислив характеристическую функцию распределения Коши, что среднее

$$\bar{\epsilon}^n = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i$$

распределено также как и случайная величина ϵ_1. Это означает, что и величина

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu$$

распределена также как и ϵ_1. То есть качество оценивания параметра μ не будет зависеть от числа проведенных наблюдений. Отметим также, что среднеквадратичный риск эмпирического среднего в данном случае бесконечен, поскольку у распределения Коши нет второго момента.

Если же для оценивания μ используется медиана, то ничего подобного не происходит. Она будет, в отличии от эмпирического среднего очень хорошей оценкой (см. теорему 3.2.3). Поэтому медиана представляет собой исключительно устойчивую оценку.

Оценка точности медианы

Контроль точности, с которой медиана оценивает параметр μ, как мы сейчас увидим, является нетривиальной задачей. Из предыдущей теоремы ясно, для оценки средне-квадратичного риска нам нужно оценить по наблюдениям величину $p(0)$. Корректное с математической точки зрения решение этой существенно выходит за рамки этого курса лекций. Тем не менее мы обсудим некоторые подходы к ее решению. Как мы знаем плотность распределения является производной от функции распределения. Поэтому мы будем аппроксимировать плотность с помощью численного дифференцирования эмпирической функции распределения $\bar{F}(x; Y^n)$.

Рассмотрим следующие величины:

$$\bar{p}_m(0; Y^n) = \frac{\bar{F}[Y_{(n/2+m)}; Y^n] - \bar{F}[Y_{(n/2-m)}; Y^n]}{Y_{(n/2+m)} - Y_{(n/2-m)}} = \frac{2m}{n[Y_{(n/2+m)} - Y_{(n/2-m)}]};$$

здесь $m = 1, 2, \ldots, n/2$ — некоторое целое число. Поэтому в качестве оценки для $p^{-1}(0)$ будем использовать одну из статистик

$$q_m(0; Y^n) = \frac{n[Y_{(n/2+m)} - Y_{(n/2-m)}]}{2m}.$$

Нас будут интересовать среднеквадратичные риски этих оценок, а именно,

$$r_m = \mathbf{E}\left[q_m(0; Y^n) - \frac{1}{p(0)}\right]^2.$$

Чтобы их вычислить, мы как обычно используем метод инверсии в сочетании с теоремой Пайка. Начнем с обращения функции распределения. При этом мы будем предполагать, что плотность распределения симметрична и является достаточно гладкой, например, имеет везде

ограниченную третью производную. Тогда из формулы Тейлора находим

$$F(x) = F(0) + F'(0)x + \frac{x^2}{2}F''(0) + \frac{x^3}{6}F^{(3)}(0) + O(x^4)$$
$$= \frac{1}{2} + xp(0) + \frac{x^3}{6}p''(0) + O(x^4).$$

Здесь мы использовали то, что $p'(0) = 0$ в силу симметрии. Из этого тождества придем у следующему выражению для обратной функции:

$$F^{-1}(y) = \frac{1}{p(0)}\left(y - \frac{1}{2}\right) - \frac{p''(0)}{6p^4(0)}\left(y - \frac{1}{2}\right)^3 + O\left(\left[y - \frac{1}{2}\right]^4\right). \qquad (3.2.9)$$

Далее из теоремы Пайка и формулы Тейлора получим (см. подробные выкладки в (3.2.6))

$$U_{(n/2+m)} - \frac{1}{2} = \frac{m}{n} + \frac{1}{2(n+1)}\sum_{i=1}^{n/2}(\zeta_i - 1)$$
$$+ \frac{1}{2(n+1)}\sum_{i=n/2+1}^{n+1}(1 - \zeta_i) + \frac{1}{n+1}\sum_{i=n/2}^{n/2+m}(\zeta_i - 1) + O\left(\frac{1}{n}\right),$$
$$U_{(n/2-m)} - \frac{1}{2} = -\frac{m}{n} + \frac{1}{2(n+1)}\sum_{i=1}^{n/2}(\zeta_i - 1)$$
$$+ \frac{1}{2(n+1)}\sum_{i=n/2+1}^{n+1}(1 - \zeta_i) + \frac{1}{n+1}\sum_{i=n/2-m}^{n/2}(1 - \zeta_i) + O\left(\frac{1}{n}\right).$$

Поэтому при $n \to \infty$

$$U_{(n/2+m)} - U_{(n/2-m)} = \frac{2m}{n} + \frac{1}{n}\sum_{i=n/2-m}^{n/2+m}(\zeta_i - 1) + O\left(\frac{1}{n}\right)$$

и

$$\left(U_{(n/2-m)} - \frac{1}{2}\right)^3 - \left(U_{(n/2+m)} - \frac{1}{2}\right)^3 = \frac{6m^2}{n^2}(1 + o(1)).$$

Отсюда и из (3.2.9) найдем

$$\begin{aligned}\frac{n[Y_{(n/2+m)}-Y_{(n/2-m)}]}{2m} &= \frac{n}{2m}\left[F_Y^{-1}(U_{(n/2+m)})-F_Y^{-1}(U_{(n/2-m)})\right]\\ &= \frac{n}{2m}\left\{\frac{1}{p(0)}\left[U_{(n/2+m)}-U_{(n/2-m)}\right]\right.\\ &\left.+\frac{p''(0)}{6p^4(0)}\left[\left(U_{(n/2-m)}-\frac{1}{2}\right)^3-\left(U_{(n/2+m)}-\frac{1}{2}\right)^3\right]\right\}\\ &= \frac{1}{p(0)}+\frac{1}{2p(0)m}\sum_{s=n/2-m}^{n/2+m}(\zeta_s-1)+\frac{p''(0)m}{2p^4(0)n}+O\left(\frac{1}{n}\right).\end{aligned}$$

Таким образом, мы находим следующую формулу для квадратичного риска

$$r_m=\left[\frac{p''(0)m}{2p^4(0)n}\right]^2+\frac{1}{2p^2(0)m}.$$

Если мы хотим иметь минимальный риск, то мы естественно должны минимизировать его по m. Дифференцируя r_m по m и приравнивая производную нулю, найдем оптимальное значение величины m

$$m^*=\left\{\frac{p^6(0)n^2}{[p''(0)]^2}\right\}^{1/3}.$$

При этом

$$r_{m^*}\asymp\frac{[p''(0)]^{2/3}}{p^3(0)n^{2/3}}.$$

Таким образом, в отличии от классической ситуации когда риск сходится со скоростью n^{-1}, здесь мы получаем более низкую скорость сходимости $n^{-2/3}$. Можно показать, что более высокой скорости невозможно достичь.

Но основная проблема заключается не в низкой скорости сходимости, а в том, что для выбора m^* нам нужно знать величину $p''(0)$. Эта величина очевидно неизвестна и оценивать ее довольно сложно, сложнее, чем $p(0)$. Вопрос о том, что надо делать в такой ситуации является базовым в современной статистике. Корректный ответ на него известен, но выходит за рамки этих лекций. Скажем только, что для оценивания $p^{-1}(0)$ мы должны строить новую оценку на основе семейства оценок

$$q_m(0;Y^n),\ m=1,2,3\ldots,n/2.$$

3.2.3 Равномерно распределенные ошибки

Теперь посмотрим как выглядит оценка максимального правдоподобия в случае когда ошибки ϵ_i в модели

$$Y_i = \mu + \epsilon_i, \quad i = 1, \dots, n$$

распределены равномерно. Это значит, что они имеют следующую плотность:

$$p(x) = \begin{cases} (2\sigma)^{-1}, & |x| \le \sigma, \\ 0, & |x| > \sigma; \end{cases}$$

здесь величина $\sigma > 0$ характеризует размах равномерного распределения.

Нетрудно проверить, что в этом случае логарифм отношения правдоподобия выглядит следующим образом:

$$\begin{aligned} L(\tilde{\mu}; Y^n) &= \sum_{i=1}^{n} \log[p(Y_i - \tilde{\mu})] \\ &= \begin{cases} -n\log(2\sigma), & \tilde{\mu} \in [Y_{(n)} - \sigma, Y_{(1)} + \sigma], \\ -\infty, & \tilde{\mu} \notin [Y_{(n)} - \sigma, Y_{(1)} + \sigma]. \end{cases} \end{aligned}$$

Из этой формулы видно, оценка максимального правдоподобия

$$\bar{\mu}(Y^n) = \arg\min_{\tilde{\mu}} L(\tilde{\mu}; Y^n)$$

не является единственной — это любое число из отрезка

$$[Y_{(n)} - \sigma, Y_{(1)} + \sigma].$$

Поэтому, как правило, в качестве оценки максимального правдоподобия берут середину этого отрезка, т.е. величину

$$\bar{\mu}(Y^n) = \frac{Y_{(1)} + Y_{(n)}}{2}. \tag{3.2.10}$$

Из теоремы Пайка и центральной предельной теоремы легко получить, что если ϵ_i *имеют равномерное распределение*, то

$$\bar{\mu}(Y^n) - \mu = \sigma \left[1 + O\left(\frac{1}{\sqrt{n}}\right)\right] \frac{\zeta_1 - \zeta_2}{n}; \tag{3.2.11}$$

здесь ζ_1, ζ_2 — независимые случайные величины, имеющие стандартное экспоненциальное распределение. Заметим, что случайная величина $\zeta_1 - \zeta_2$ имеет стандартное распределение Лапласа.

Из (3.2.11) легко найти, что

$$\mathbf{E}[\bar{\mu}(Y^n) - \mu]^2 = \left[1 + O\left(\frac{1}{\sqrt{n}}\right)\right]\frac{4\sigma^2}{n^2}.$$

Мы видим, что в данном случае среднеквадратичный риск убывает со скоростью $1/n^2$. Напомним, что как мы видели ранее, если для оценки μ используются эмпирические средние или медиана, то соответствующие скорости будут иметь порядок $1/n$.

Посмотрим теперь на устойчивость рассматриваемой оценки по отношению к выбору модели. Предположим, что ϵ_i имеют стандартное распределение Лапласа, а параметр μ оценивается с помощью (3.2.10). В этом случае легко доказать (см. теорему 1.3.8), что

$$\begin{aligned} Y_{(1)} &= \log(n) - \log(\zeta_1) + \mu + O\left(\frac{1}{\sqrt{n}}\right), \\ Y_{(n)} &= -\log(n) + \log(\zeta_2) + \mu + O\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

где ζ_1, ζ_2 — независимые случайные величины, имеющие стандартное экспоненциальное распределение. Поэтому

$$\bar{\mu}(Y^n) - \mu = \log(\zeta_2) - \log(\zeta_1) + O\left(\frac{1}{\sqrt{n}}\right).$$

То есть, с ростом числа наблюдений n оценка не будет приближаться ни в каком вероятностном смысле к оцениваемому параметру. Это значит, что эта оценка является неустойчивой по отношению к выбору модели; она может быть как очень хорошей, так и очень плохой. Поэтому она практически никогда не используется для оценивания параметра сдвига.

Глава 4

Байесовское оценивание

4.1 Сравнение статистических оценок

Будем предполагать, что статистические данные $Y^n = \{Y_1, \ldots, Y_n\}$, которыми располагает статистик, могут быть описаны семейством вероятностных плотностей $p_\theta(y)$, зависящих от конечномерного параметра $\theta \in \Theta \subset \mathbb{R}^p$. Это предположение означает, найдется такой вектор $\theta \in \Theta$, что

$$\begin{aligned} &\mathbf{P}\Big\{Y_i \in [y_i, y_i + \Delta_i],\ i = 1, \ldots, n\Big\} \\ &\qquad = p_\theta(y_1, \ldots, y_n) \cdot \Delta_1 \cdot \ldots \cdot \Delta_n (1 + o(1)) \end{aligned} \tag{4.1.1}$$

когда $\max_{i=1,\ldots,n} |\Delta_i| \to 0$.

Основная задача статистики состоит в том, чтобы по наблюдениям Y^n оценить параметр θ. Термин оценить означает найти некоторую функцию

$$\bar{\theta}(y_1, \ldots, y_n) :\ \mathbb{R}^n \to \mathbb{R}^p$$

такую, чтобы вектор $\bar{\theta}(Y_1, \ldots, Y_n)$ был бы как можно ближе к вектору $\theta \in \Theta$.

В качестве оценки можно взять, например, оценку максимального правдоподобия

$$\bar{\theta}(Y^n) = \arg\max_{\tilde{\theta} \in \Theta} p(\tilde{\theta}; Y^n),$$

а можно просто взять некоторый произвольный вектор $\theta_\circ$ из Θ.

Понятно, что для того, чтобы сказать насколько одна оценка находится ближе к оцениваемому параметру θ, чем другая, нам нужно определить расстояние между элементами в Θ. В статистике это расстояние

задается так называемой *функцией потерь*

$$l(\hat{\theta}, \theta) : \ \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+.$$

Смысл этой функции — приписать паре {*оценка, параметр*} положительное число, которое показывало бы насколько оценка далека от оцениваемого объекта. Выбор этой функции основан на физическом смысле, который имеет оцениваемый параметр и вся задача оценивания в целом.

Если Θ — дискретное множество, т. е. множество состоящее из конечного набора векторов, то естественно взять

$$l(\hat{\theta}, \theta) = \mathbf{1}\{\hat{\theta} \neq \theta\}.$$

В случае когда Θ, например, все пространство $\mathbb{R}^p$, то часто используется квадратичная функция потерь

$$l(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2;$$

здесь $\|\cdot\|$ — стандартная евклидова норма в $\mathbb{R}^p$.

Очевидно, что потери измеряемые при помощи функции потерь зависят от наблюдений, что не очень удобно в смысле сравнения оценок. Поэтому, как правило, качество оценки $\hat{\theta}(Y^n)$ измеряется *риском*

$$R(\hat{\theta}, \theta) = \mathbf{E} l\big[\hat{\theta}(Y^n), \theta\big] = \int_{\mathbb{R}^n} l\big[\hat{\theta}(y_1, \dots, y_n), \theta\big] p(y_1, \dots, y_n; \theta)\, dy_1 \cdots dy_n$$

который представляет собой ни что иное, как средние потери. Риск зависит от двух объектов:

- оценки, которая является функцией $\mathbb{R}^n \to \mathbb{R}^p$,
- оцениваемого параметра, который принадлежит множеству $\Theta \subset \mathbb{R}^p$.

Наша цель — найти наилучшую оценку, т.е. оценку, которая имела бы минимальный риск. Практически очевидно, что эта задача не имеет решения, поскольку невозможно сказать, сравнивая риски, какая из двух оценок $\hat{\theta}_1$ или $\hat{\theta}_2$ лучше. Дело в том, что для того, чтобы сравнить оценки, нужно сравнить две функции $f_1(\theta) = R(\hat{\theta}_1, \theta)$ и $f_2(\theta) = R(\hat{\theta}_2, \theta)$ при всех $\theta \in \Theta$. Понятно, что сделать это, как правило, нельзя поскольку сравнивать можно только числа, но не функции. Поэтому самое естественное решение приписать риску некоторое число. Проще всего это сделать с помощью линейного функционала

$$r_\pi(\hat{\theta}) = \int_{\mathbb{R}^p} \pi(\theta) R(\hat{\theta}, \theta)\, d\theta_1 \cdots d\theta_p$$

здесь $\pi(\theta) : \mathbb{R}^p \to \mathbb{R}^+$ — некоторая неотрицательная функция, которая обращается в нуль при $\theta \notin \Theta$. Ясно, что без ограничения общности можно считать, что

$$\int_{\mathbb{R}^p} \pi(\theta)\, d\theta = 1.$$

Вероятностная интерпретация риска $r^\pi(\hat\theta)$ очевидна: предполагается, что оцениваемый параметр θ является случайной величиной с плотностью распределения $\pi(\theta)$ и риск вычисляется как усредненные потери по распределению наблюдений и по распределению неизвестного параметра.

Поскольку сравнивать числа можно, то, естественно, возможно определить оценку

$$\hat\theta_\pi(Y^n) = \arg\min_{\bar\theta} r_\pi(\bar\theta),$$

где min вычисляется по *всем оценкам параметра* θ, т.е. по всем функциям $\mathbb{R}^n \to \mathbb{R}^p$. Эта оценка называется *байесовской оценкой.*

Хотя на первый взгляд кажется, что $\hat\theta_\pi$ трудно вычислить, в действительности, это не так.

Теорема 4.1.1 *Если существует*

$$\hat\theta_\pi(Y^n) = \arg\min_{\tilde\theta\in\mathbb{R}^p} \int_{\mathbb{R}^p} \pi(\theta) l(\tilde\theta,\theta) p(Y_1,\dots,Y_n;\theta)\, d\theta, \tag{4.1.2}$$

то $\hat\theta_\pi(Y^n)$ *— байесовская оценка.*

Доказательство. Утверждение этой теоремы эквивалентно утверждению о том, что риск любой оценки не меньше риска оценки $\hat\theta_\pi(Y^n)$. Заметим, что меняя порядок интегрирования, получим, что для любой оценки $\bar\theta(Y^n)$ справедливо неравенство

$$\begin{aligned}
r_\pi(\bar\theta) &= \int_{\mathbb{R}^p} \pi(\theta) \int_{\mathbb{R}^n} l[\bar\theta(y),\theta] p(y;\theta)\, dy\, d\theta \\
&= \int_{\mathbb{R}^n} \left\{ \int_{\mathbb{R}^p} \pi(\theta) l[\bar\theta(y),\theta] p(y;\theta)\, d\theta \right\} dy \\
&\geq \int_{\mathbb{R}^n} \min_{\tilde\theta} \left\{ \int_{\mathbb{R}^p} \pi(\theta) l[\tilde\theta,\theta] p(y;\theta)\, d\theta \right\} dy \\
&= \int_{\mathbb{R}^n} \left\{ \int_{\mathbb{R}^p} \pi(\theta) l[\hat\theta_\pi(y),\theta] p(y;\theta)\, d\theta \right\} dy = r_\pi(\hat\theta_\pi). \quad \blacksquare
\end{aligned}$$

Обозначим для краткости

$$Z(Y^n) = \int_{\mathbb{R}^p} \pi(\theta) p(Y^n; \theta)\, d\theta.$$

Тогда

$$q(\theta; Y^n) = \frac{\pi(\theta) p(Y^n; \theta)}{Z(Y^n)}$$

называется *апостериорной плотностью* распределения параметра θ, а

$$\rho(\bar{\theta}; Y^n) = \int_{\mathbb{R}^p} l\big[\bar{\theta}(Y^n), \theta\big] q(\theta; Y^n)\, d\theta$$

апостериорным риском оценки $\bar{\theta}(Y^n)$.

Апостериорный риск предсказывает (оценивает) риск, который будет иметь оценка $\bar{\theta}(Y^n)$. Байсовская оценка это — оценка с минимальным апостериорным риском.

Формула (4.1.2) дает только формальный рецепт вычисления наилучшей оценки. Практически ее использовать, как правило, трудно. Далее мы будем заниматься тем, чтобы на основе этой формулы получить оценки, не требующие решения оптимизационных задач.

Начнем со случая когда функция потерь является квадратичной

$$l(\tilde{\theta}, \theta) = \|\tilde{\theta} - \theta\|^2 = \sum_{i=1}^{p} (\tilde{\theta}_i - \theta_i)^2.$$

Тогда для того, чтобы вычислить

$$\hat{\theta}_\pi(Y^n) = \arg\min_{\tilde{\theta}} \int_{\mathbb{R}^p} \|\tilde{\theta} - \theta\|^2 \pi(\theta) p(Y^n; \theta)\, d\theta$$

достаточно решить уравнения

$$\int_{\mathbb{R}^p} (\tilde{\theta}_i - \theta_i) \pi(\theta) p(Y^n; \theta)\, d\theta = 0,\ i = 1, \ldots, p.$$

Отсюда сразу же находим

$$\hat{\theta}_i(Y^n) = \frac{1}{Z(Y^n)} \int_{\mathbb{R}^p} \theta_i \pi(\theta) p(Y^n; \theta)\, d\theta.$$

Иными словами, при квадратичных потерях байсовская оценка является апостериорным средним. Заметим, что это тоже формальное решение задачи, поскольку подсчет многомерных интегралов является довольно сложной задачей. Поэтому задача упрощения этой оценки для того, чтобы избежать многомерного численного интегрирования, играет очень важную роль в статистике.

4.2 Неравенство Ван Триса

Если наша вероятностная модель не является гауссовской, то написать уравнения для наилучшей оценки и вычислить ее риск может оказаться достаточно трудно. Тем не менее, для баейсовского риска есть очень хорошая граница снизу.

Очевидно, что за качество баейсовской оценки $\hat{\theta}_\pi$ отвечают две плотности

- плотность распределения наблюдений $p(x;\theta)$,
- априорная плотность распределения неизвестного параметра $\pi(\theta)$.

С этими плотностями связаны две информационные матрицы Фишера

$$I_p = \mathbf{E}\left(\frac{\partial}{\partial\theta}\log[p(Y^n;\theta)]\right)\left(\frac{\partial}{\partial\theta}\log[p(Y^n;\theta)]\right)^\top,$$
$$I_\pi = \mathbf{E}\left(\frac{\partial}{\partial\theta}\log[\pi(\theta)]\right)\left(\frac{\partial}{\partial\theta}\log[\pi(\theta)]\right)^\top;$$

здесь $\mathbf{E}$ – усреднение по совместной плотности распределения вектора (Y^n,θ), т.е

$$\mathbf{E}f(Y^n,\theta) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^p} f(y,u)p(y;u)\pi(u)\,dy\,du.$$

Вероятностный смысл информационного количества Фишера

$$I(\theta) = \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta}\log[p(Y^n;\theta)]\right)^2,$$

где $\mathbf{E}_\theta$ означает усреднение при фиксированном параметре θ, достаточно прозрачен. Пусть для простоты θ — одномерный параметр. Тогда

$$I(\theta) = \int_{\mathbb{R}^n}\frac{[p'_\theta(x;\theta)]^2}{p(x;\theta)}\,dx;$$

здесь

$$p'_\theta(x) = \frac{dp(x;\theta)}{d\theta}.$$

С другой стороны, для расстояния Кульбака-Лейблера найдем, пользуясь формулой

$$\log(1+x) \approx x - \frac{x^2}{2},$$

что при малых h

$$K[p(\cdot;\theta,p(\cdot;\theta+h)] \approx \frac{1}{2}\int_{\mathbb{R}^n}\frac{[p(x;\theta+h)-p(x;\theta)]^2}{p(x;\theta)}\,dx \approx \frac{h^2 I(\theta)}{2}.$$

То есть, информационное количество Фишера связывает локально расстояние Кульбака-Лейблера и стандартное евклидово расстояние.

Теорема 4.2.1 *Предположим, что существуют и конечны фишеровские информации I_p, I_π и $I_p + I_\pi > 0$. Тогда для любой оценки $\bar\theta$ выполнено неравенство*

$$\mathbf{E}(\bar\theta-\theta)(\bar\theta-\theta)^\top \geq (I_p+I_\pi)^{-1}. \tag{4.2.3}$$

Доказательство. Ограничимся для простоты и наглядности случаем $\theta\in\mathbb{R}$. При этом будем предполагать для простоты, что $\mathbf{E}|\theta|^2<\infty$ и что при любом фиксированном $y\in\mathbb{R}^n$

$$\lim_{|\theta|\to\infty}|\theta|\pi(\theta)p(y;\theta)=0.$$

Основная идея в доказательстве теоремы 4.2.1 - применить неравенство Коши-Буняковского

$$\big[\mathbf{E}a(Y^n,\theta)\cdot b(Y^n,\theta)\big]^2 \leq \mathbf{E}a^2(Y^n,\theta)\cdot\mathbf{E}b^2(Y^n,\theta), \tag{4.2.4}$$

которое справедливо для любых случайных величин $a(Y^n,\theta)$ и $b(Y^n,\theta)$ с ограниченными вторыми моментами.

Возьмем

$$a(Y^n,\theta)\stackrel{\text{def}}{=}\bar\theta(Y^n)-\theta,\quad b(Y^n,\theta)\stackrel{\text{def}}{=}\frac{d}{d\theta}\log[p(Y^n;\theta)\pi(\theta)].$$

Тогда, интегрируя по частям

$$\begin{aligned}\mathbf{E}a(Y^n,\theta)b(Y^n,\theta) &= \int_{\mathbb{R}^n}\left\{\int_{\mathbb{R}^1}[\bar\theta(y)-\theta]\,d_\theta[p(y;\theta)\pi(\theta)]\right\}dy\\ &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^1}p(y;\theta)\pi(\theta)\,dy\,d\theta=1.\end{aligned} \tag{4.2.5}$$

Здесь мы неявно использовали, что

$$\lim_{\theta\to\pm\infty}\theta p(y;\theta)\pi(\theta)=0.$$

Далее

$$\begin{aligned}\mathbf{E}b^2(Y,\theta) &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^1}\left[\frac{d}{d\theta}\log[p(y;\theta)\pi(\theta)]\right]^2 p(y;\theta)\pi(\theta)\,dy\,d\theta \\ &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^1}\left[\frac{1}{p(y;\theta)}\frac{dp(y;\theta)}{d\theta}+\frac{1}{\pi(\theta)}\frac{d\pi(\theta)}{d\theta}\right]^2 p(y;\theta)\pi(\theta)\,dy\,d\theta \\ &= I_p + I_\pi + 2\int_{\mathbb{R}^n}\int_{\mathbb{R}^1}\frac{d\pi(\theta)}{d\theta}\frac{dp(y;\theta)}{d\theta}dy\,d\theta \\ &= I_p + I_\pi + 2\int_{\mathbb{R}^1}\left\{\frac{d}{d\theta}\int_{\mathbb{R}^n}p(y;\theta)\,dy\right\}\frac{d\pi(\theta)}{d\theta}\,d\theta = I_p + I_\pi.\end{aligned}$$

Это равенство, (4.2.4) и (4.2.5) завершают доказательство теоремы. ■

Применим неравенство Ван Триса к модели оценивания параметра сдвига μ по наблюдениям

$$Y_i = \mu + \epsilon_i,\ i = 1,\dots,n,$$

где ϵ_i — независимые гауссовские случайные величины с $\mathbf{E}\epsilon_i = 0$ и $\mathbf{E}\epsilon_i^2 = \sigma^2$.

Предположим, что параметр μ является гауссовской случайной величиной с нулевым средним и дисперсией $\mathbf{E}\mu^2 = \Sigma^2$. Это значит, что априорная плотность распредления μ имеет вид

$$\pi(\mu) = \exp\left\{-\frac{\mu^2}{2\Sigma^2} - \frac{1}{2}\log(2\pi\Sigma^2)\right\},$$

а плотность распределения наблюдений определяется как

$$p_\mu(y_1,\dots,y_n) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\mu)^2 - \frac{n}{2}\log(2\pi\sigma^2)\right\}.$$

Подсчет информационных количеств очень прост в данном случае. Имеем

$$I_\pi = \mathbf{E}\left[\frac{d}{d\mu}\log[\pi(\mu)]\right]^2 = \mathbf{E}\left[-\frac{\mu}{\Sigma^2}\right]^2 = \frac{1}{\Sigma^2}$$

и совершенно аналогично

$$I_p = \mathbf{E}\left[\frac{d}{d\mu}\log[p_\mu(Y^n)]\right]^2 = \mathbf{E}\left[-\frac{1}{\sigma^2}\sum_{i=1}^n(Y_i-\mu)\right]^2 = \frac{n}{\sigma^2}.$$

Поэтому в силу неравенства Ван Триса для любой оценки $\tilde{\mu}(Y^n)$ находим

$$\mathbf{E}(\tilde{\mu}-\mu)^2 \geq [I_\pi + I_p]^{-1} = \frac{\sigma^2\Sigma^2}{\sigma^2 + n\Sigma^2}.$$

Рассмотрим теперь линейную оценку

$$\bar{\mu}_\pi(Y^n) = \frac{\Sigma^2}{\sigma^2 + n\Sigma^2} \sum_{i=1}^{n} Y_i.$$

и сосчитаем ее средне-квадратичный риск. Находим

$$\begin{aligned}\mathbf{E}(\bar{\mu}_\pi - \mu)^2 =& \mathbf{E}\left[\mu - \frac{\mu n\Sigma^2}{\sigma^2 + n\Sigma^2} + \frac{\Sigma^2}{\sigma^2 + n\Sigma^2} \sum_{i=1}^{n} \epsilon_i\right]^2 \\ =& \frac{\Sigma^2\sigma^4}{(\sigma^2 + n\Sigma^2)^2} + \frac{\Sigma^4\sigma^2 n}{(\sigma^2 + n\Sigma^2)^2} = \frac{\sigma^2\Sigma^2}{\sigma^2 + n\Sigma^2}.\end{aligned}$$

Таким образом, верхняя и нижняя границы для байесовского риска совпадают и оценка $\bar{\mu}_\pi(Y^n)$ является байесовской оценкой. Это факт не случаен. Далее мы увидим, что в гауссовских моделях байесовские оценки являются всегда линейными.

Сейчас же посмотрим, что будет происходить с байессовской оценкой и риском при большом числе наблюдений, т.е. при $n \to \infty$. Вспомним, что эмпирическое среднее $\bar{Y}^n$ является оценкой максимального правдоподобия параметра μ.

Асимптотические свойства байесовской оценки и риска.

- $$\frac{\bar{\mu}_\pi(Y^n) - \bar{Y}^n}{\bar{Y}^n} = O\left(\frac{\sigma^2}{n\Sigma^2}\right),$$
- $$\frac{\mathbf{E}[\bar{\mu}_\pi(Y^n) - \mu]^2 - \mathbf{E}[\bar{Y}^n - \mu]^2}{\mathbf{E}[\bar{Y}^n - \mu]^2} = O\left(\frac{\sigma^2}{n\Sigma^2}\right).$$

Таким образом видим, что оценка максимального правдоподобия является асимптотической аппроксимацией байесовских оценок. Это одно из принципиально важных свойств метода максимального правдоподобия. Асимптотическая оптимальность этого метода состоит в том, что он

- не зависит от априорной плотности $\pi(\cdot)$;
- при $n \to \infty$ приближается к наилучшей оценке параметра, построенной при заданной плотности $\pi(\cdot)$.

Заметим еще, что оценку максимального правдоподобия можно рассматривать как байесовскую оценку при $\Sigma^2 = \infty$, то есть при нулевой фишеровской информации об оцениваемом параметре.

4.3 Байесовские оценки в гауссовских моделях

Те простые факты, которые мы установили для простейшей модели оценивания параметра сдвига, можно обобщить на существенно более широкий класс гауссовских моделей.

Определение 4 *Случайный вектор* $Y^n = (Y_1, \dots, Y_n)^\top$ *имеет гауссовское распределения со средним* $m = (m_1, \dots, m_n)^\top$ *и корреляционной матрицей* $B = \{B_{ij},\, i, j = 1, \dots, n\}$*, если его плотность распределения имеет вид*

$$p(y; m, B) = \exp\left\{-\frac{\langle B^{-1}(y-m), y-m\rangle}{2} - \frac{n}{2}\log[2\pi\det(B)]\right\}; \quad (4.3.6)$$

здесь $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$*,* $\langle\cdot,\cdot\rangle$ *— скалярное произведение в* $\mathbb{R}^n$*,* B^{-1} *— матрица обратная к* B*,* $\det(B)$ *— детерминант матрицы* B*.*

Вероятностный смысл параметров гауссовского распределения

- $$m_i = \mathbf{E}Y_i = \int_{\mathbb{R}^n} y_i p(y_1, \dots, y_n)\, dy_1 \cdots dy_n,$$
- $$\begin{aligned} B_{i,j} =& \mathbf{E}(Y_i - m_i)(Y_j - m_j) \\ =& \int_{\mathbb{R}^n} (y_i - m_i)(y_j - m_j) p(y_1, \dots, y_n)\, dy_1 \cdots dy_n. \end{aligned}$$

Напомним, что матрица $A = \{A_{i,j},\, i, j = 1, \dots, n\}$ называется неотрицательно (положительно) определенной если для любых чисел t_i, $i = 1, \dots, n$ выполнено неравенство

$$\sum_{i,j=1}^{n} A_{i,j} t_i t_j \geq (>) 0.$$

Теорема 4.3.1 *Корреляционная матрица* $B = \mathbf{E}(Y-m)(Y-m)^\top$ *неотрицательно определена.*

Доказательство. Рассмотрим случайную величину

$$\xi = \sum_{i=1}^{n} t_i (Y_i - m_i).$$

и заметим, что

$$0 \le \mathbf{E}\xi^2 = \mathbf{E}\sum_{i=1}^n\sum_{j=1}^n t_i t_j (Y_i - m_i)(Y_j - m_j)$$
$$= \sum_{i=1}^n\sum_{j=1}^n t_i t_j \mathbf{E}(Y_i - m_i)(Y_j - m_j) = \sum_{i=1}^n\sum_{j=1}^n t_i t_j B_{ij}. \quad \blacksquare$$

Хорошо известно, что если B — положительно определенная $n \times n$-матрица, то существуют собственные векторы $\phi_i \in \mathbb{R}^n$ и собственные числа $\lambda_i \in \mathbb{R}^+$ такие, что

$$B\phi_i = \lambda_i \phi_i,\ i = 1, \dots, n$$

и справедлив следующий результат:

Теорема 4.3.2

$$B = \sum_{i=1}^n \lambda_i \phi_i \phi_i^\top,\ B^{-1} = \sum_{i=1}^n \lambda_i^{-1} \phi_i \phi_i^\top,\ \det(B) = \prod_{i=1}^n \lambda_i. \tag{4.3.7}$$

Вероятностная версия этого результата имеет следующий вид:

Теорема 4.3.3 *Пусть $\{\xi_1, \dots \xi_n\}$ — независимые стандартные гауссовские случайные величины, а Y^n — гауссовский вектор с плотностью из* (4.3.6)*. Тогда*

$$Y^n = m + \sum_{i=1}^n \sqrt{\lambda_i}\xi_i \phi_i, \tag{4.3.8}$$

$$p(y; m, B) = \exp\left\{ -\frac{1}{2}\sum_{i=1}^n \frac{\langle y - m, \phi_i\rangle^2}{\lambda_i} - \frac{n}{2}\sum_{i=1}^n \log(\lambda_i) \right.$$
$$\left. - \frac{n\log(2\pi)}{2} \right\}; \tag{4.3.9}$$

здесь λ_i и ϕ_i — собственные числа и собственные векторы ковариационной матрицы B.

Доказательство. Достаточно проверить, что гауссовский вектор Y^n из (4.3.8) имеет ковариационную матрицу B. Действительно,

$$\mathbf{E}(Y^n - m)(Y^n - m)^\top = \mathbf{E}\sum_{i,j=1}^n \sqrt{\lambda_i \lambda_j}\phi_i \phi_j^\top \xi_i \xi_j$$
$$= \sum_{i,j=1}^n \sqrt{\lambda_i\lambda_j}\phi_i\phi_j^\top \mathbf{E}\xi_i\xi_j = \sum_{i=1}^n \lambda_i \phi_i \phi_i^\top = B.$$

Формула (4.3.9) вытекает непосредственно из (4.3.8) в силу того, что случайные величины $\sqrt{\lambda_i}\xi_i$ — независимые и гауссовские. ■

Теперь мы обобщим задачу оценивания среднего гауссовского вектора. Более общая задача формулируется следующим образом. Предположим, что наблюдается гауссовский вектор $Y^n = (Y_1, \dots, Y_n)^\top$ и по этим наблюдениям нужно оценить случайный вектор $\theta^p = (\theta_1, \dots, \theta_p)^\top \in \mathbb{R}^p$.

Теорема 4.3.4 *Если совместное распределение случайных векторов $Y^n = \{Y_1, \dots, Y_n\}$ и $\theta^p = \{\theta_1, \dots, \theta_p\}$ является гауссовским с нулевым средним и с плотностью $p(y,t)$, $y \in \mathbb{R}^n$, $t \in \mathbb{R}^p$, то байесовская оценка*

$$\bar{\theta}_i(Y^n) = \int_{\mathbb{R}^p} t_i p(Y^n;t)\,dt_1\cdots dt_p \Big/ \int_{\mathbb{R}^p} p(Y^n;t)\,dt_1\cdots dt_p \tag{4.3.10}$$

является линейной по наблюдениям Y^n, т.е.

$$\bar{\theta}_i(Y^n) = \sum_{s=1}^n K_{is} Y_s, \tag{4.3.11}$$

где ядро K_{is}, $i = 1, \dots, p$ $s = 1, \dots, n$ удовлетворяет уравнению Винера-Хопфа

$$\sum_{s=1}^n K_{is}\mathbf{E}Y_s Y_k = \mathbf{E}Y_k\theta_i,\ k = 1\ \dots, n. \tag{4.3.12}$$

При этом апостериорные риски

$$\rho_i = \int_{\mathbb{R}^p} [t_i - \bar{\theta}_i(Y^n)]^2 p(Y^n;t)\,dt_1\cdots dt_p \Big/ \int_{\mathbb{R}^p} p(Y^n;t)\,dt_1\cdots dt_p \tag{4.3.13}$$

не зависят от наблюдений Y^n и вычисляются как

$$\rho_i = \mathbf{E}\theta_i^2 - \sum_{s=1}^n K_{is}\mathbf{E}\theta_i Y_s.$$

Доказательство. Обозначим B ковариационную матрицу гауссовского вектора $\{Y^n, \theta_i\} \in \mathbb{R}^{n+1}$, а B^{-1} ее обратную. Тогда из определения гауссовской плотности получим

$$\begin{aligned}\log p(y;t) &= -\frac{\left\langle B^{-1}\cdot\{y,t\},\{y,t\}\right\rangle}{2} - \frac{n}{2}\log[2\pi\det(B)] \\ &= -\frac{t^2 B^{-1}_{n+1,n+1}}{2} - t\sum_{s=1}^n B^{-1}_{s,n+1}y_s + C(y),\end{aligned} \tag{4.3.14}$$

где $C(y)$ — некоторая функция Y, независящая от t. Эту формулу легко понять в силу того, что s - ая компонента вектора $B^{-1}\cdot\{y,t\}$ имеет вид

$$\left[[B^{-1}\cdot\{y,t\}\right]_s = C_s(y) + B^{-1}_{s,n+1}t;$$

здесь $C_s(y)$ — некоторая функция от y.

Далее, выделяя в правой части (4.3.14) полный квадрат, найдем

$$\log p(y,t) = -\frac{1}{2}\left(\sqrt{B^{-1}_{n+1,n+1}}t + \sum_{s=1}^{n} B^{-1}_{s,n+1}y_s \Big/ \sqrt{B^{-1}_{n+1,n+1}}\right)^2 + C(y).$$

Поэтому, делая замену переменных

$$u = \sqrt{B^{-1}_{n+1,n+1}}t + \sum_{s=1}^{n} B^{-1}_{s,n+1}y_s \Big/ \sqrt{B^{-1}_{n+1,n+1}} \tag{4.3.15}$$

или, что эквивалентно,

$$t = \frac{u}{\sqrt{B^{-1}_{n+1,n+1}}} - \frac{1}{B^{-1}_{n+1,n+1}}\sum_{s=1}^{n} B^{-1}_{s,n+1}y_s$$

при вычислении интегралов, входящих в правую часть (4.3.10), убеждаемся в линейности $\bar{\theta}_i(Y)$ по Y. Точнее получаем (4.3.11)

$$\bar{\theta}_i(Y^n) = -\sum_{s=1}^{n}\frac{B^{-1}_{s,n+1}}{B^{-1}_{n+1,n+1}}Y_s = \sum_{i=1}^{n} K_{is}Y_s.$$

Поскольку в силу теоремы 4.1.1 оценка $\bar{\theta}_i(Y)$ минимизирует байесовский риск, то

$$K_{is} = \arg\min_{\tilde{K}\in\mathbb{R}^n} \mathbf{E}\left[\theta_i - \sum_{s=1}^{n}\tilde{K}_sY_s\right]^2.$$

Это означает что K_{is} удовлетворяют уравнению

$$\frac{\partial}{\partial\tilde{K}_k}\mathbf{E}\left(\theta_i - \sum_{s=1}^{n}\tilde{K}_sY_s\right)^2\Big|_{\tilde{K}_s=K_{is}} = 0,$$

которое очевидно эквивалентно (4.3.12).

Независимость апостериорного риска ρ_i от наблюдений вытекает из подсчета интегралов, входящих в его определение (4.3.13). При этом используется замена переменных из (4.3.15). ■

4.4 Статистика гауссовских стационарных последовательностей

До сих пор мы в основном рассматривали статистические выводы в задачах, связанных с независимыми случайными величинами. Это безусловно, важная область теории статистических выводов, но это, конечно, частный случай зависимых наблюдений. Простейшей математической моделью, описывающей зависимости в статистических данных являются стационарные последовательности.

4.4.1 Представления стационарных последовательностей

Далее, далее не оговаривая каждый раз этого особо, будем считать, что все случайные величины в этом разделе являются гауссовскими.

Определение 5 *Последовательность гауссовских случайных величин* Y_k, $k = 0, \pm1, \pm2, \ldots$ *с нулевым средним* $\mathbf{E}Y_k = 0$ *называется стационарной, если равенство*

$$\mathbf{E}Y_{s+k}Y_s = \mathbf{E}Y_0Y_k,$$

выполнено для всех $k, s = 0, \pm1, \pm2, \ldots$.

Определение 6 *Функция*

$$f_Y(\lambda) = \sum_{k=-\infty}^{\infty} \exp(2\pi i\lambda k)\mathbf{E}Y_0Y_k, \quad \lambda \in [-1/2, 1/2],$$

называется спектральной плотностью стационарной последовательности Y_k.

Последовательность независимых гауссовских случайных величин ξ_k с $\mathbf{E}\xi_k = 0$ и $\mathbf{E}\xi_k^2 = 1$ называют часто стандартным *белым шумом* с дискретным временем.

Заметим, что поскольку $\mathbf{E}\xi_l\xi_k = 0$ при $k \neq l$, то спектральная плотность белого шума $f_\xi(\lambda) = 1$ при всех $\lambda \in [-1/2, 1/2]$. Именно это свойство объясняет почему шум называется белым.

С белым гауссовским шумом с дискретным временем тесно связано пространство $l_2(-\infty, \infty)$, которое содержит все последовательности φ_k, $k = 0, \pm1, \ldots$ такие, что

$$\|\varphi\|^2 \stackrel{\text{def}}{=} \sum_{k=-\infty}^{\infty} |\varphi_k|^2 < \infty.$$

Это гильбертово пространство со скалярным произведением

$$\langle \varphi, \psi \rangle \stackrel{\text{def}}{=} \sum_{k=-\infty}^{\infty} \varphi_k \psi_k^*$$

Теорема 4.4.1 *Для любой последовательности $\varphi \in l_2(-\infty, \infty)$ существует гауссовская случайная величина*

$$\langle \varphi, \xi \rangle = \sum_{k=-\infty}^{\infty} \varphi_k \xi_k.$$

Причем для любых $\varphi, \psi \in l_2(-\infty, \infty)$

$$\mathbf{E}\langle \varphi, \xi \rangle \langle \psi, \xi \rangle = \langle \varphi, \psi^* \rangle. \tag{4.4.16}$$

Доказательство. Для любой последовательности $\varphi \in l_2(-\infty, \infty)$ определим

$$\varphi_k^N = \begin{cases} \varphi_k, & |k| \le N, \\ 0, & |k| > N. \end{cases}$$

Тогда утверждение теоремы эквивалентно существованию предела последовательности случайных величин

$$\eta_N = \langle \varphi^N, \xi \rangle.$$

Заметим, что эта случайная величина имеет плотность

$$p_{\eta^N}(x) = \exp\left\{ -\frac{x^2}{2\|\varphi^N\|^2} - \frac{1}{2} \log\left[2\pi \|\varphi^N\|^2\right] \right\}$$

Тогда расстояние Кульбака-Лейблера между плотностями $p_{\eta^N}(x)$ и $p_{\eta^M}(x)$ вычисляется следующим образом:

$$\begin{aligned} K\left(p_{\eta^N}, p_{\eta^M}\right) =& \frac{1}{\sqrt{2\pi\|\varphi^N\|^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/(2\|\varphi^N\|^2)} \\ & \times \left[-\frac{x^2}{2} \left(\frac{1}{\|\varphi^N\|^2} - \frac{1}{\|\varphi^M\|^2} \right) - \frac{1}{2} \log \frac{\|\varphi^N\|^2}{\|\varphi^M\|^2} \right] dx \\ =& \frac{1}{2} \left(\frac{\|\varphi^M\|^2}{\|\varphi^N\|^2} - 1 \right) + \frac{1}{2} \log \frac{\|\varphi^M\|^2}{\|\varphi^N\|^2}. \end{aligned}$$

Отсюда видно, что при $N, M \to \infty$ расстояние Кульбака-Лейблера стремиться к нулю и, следовательно, по неравенству Пинскера (см. теорему 2.2.1) расстояние по вариации между плотностями $p_{\eta^N}(x)$ и $p_{\eta^M}(x)$

также стремится к нулю. Это доказывает существование предельной плотности распределения, что эквивалентно существованию предельной случайной величины.

Тождество (4.4.23) выполняется для любых последовательностей ϕ^N и ψ^M.

$$\mathbf{E}\langle\varphi^N,\xi\rangle\langle\psi^M,\xi\rangle = \langle\varphi^N,\psi^M\rangle.$$

Поэтому предельный переход этом тождестве сначала при $N\to\infty$, а затем при $M\to\infty$ завершает доказательство теоремы. ■

Отметим, что если рассматривать ξ_k как детерминированную последовательность, то для того, чтобы скалярное произведение $\langle\varphi,\xi\rangle$ существовало для любого $\varphi\in l_2(-\infty,\infty)$ нужно, чтобы и $\xi\in l_2(-\infty,\infty)$. Этот пример показывает, что для случайных последовательностей могут существовать объекты, которые не существуют для детерминированных.

Определение стационарной последовательности является в некоторой степени формальным, поскольку из него не видно непосредственно как можно генерировать стационарные последовательности. Слово *генерировать* означает создать стационарную гауссовскую последовательность из белого гауссовского шума, поскольку только белый шум, как правило, имеется в нашем распоряжении. Самый простой метод описан в следующей теореме.

Теорема 4.4.2 *Пусть ξ_k, $k=0,\pm1,\pm2,\ldots$ — стандартный, белый гауссовский шум. Тогда стационарная гауссовская последовательность Y_k со спектральной плотностью $f_Y(\lambda)$ может быть представлена в виде*

$$Y_k=\sum_{s=-\infty}^{\infty}h_{k-s}\xi_s, \tag{4.4.17}$$

где

$$h_m=\int_{-1/2}^{1/2}\sqrt{f_Y(\lambda)}\mathrm{e}^{2\pi\mathrm{i}\lambda m}\,d\lambda. \tag{4.4.18}$$

Доказательство. Легко видеть, что корреляционная функция последовательности Y_k из (4.4.17) вычисляется следующим образом:

$$\begin{aligned}\mathbf{E}Y_kY_l=\mathbf{E}\sum_{s=-\infty}^{\infty}h_{k-s}\xi_s\sum_{m=-\infty}^{\infty}h_{l-m}\xi_m&=\sum_{s=-\infty}^{\infty}h_{k-s}h_{l-s}\\&=\sum_{s=-\infty}^{\infty}h_{k-l+s}h_s.\end{aligned} \tag{4.4.19}$$

Отсюда сразу же находим, что спектральная плотность этой последовательности вычисляется как

$$f(\lambda) = \sum_{k=-\infty}^{\infty} \mathrm{e}^{2\pi \mathrm{i}\lambda k} \sum_{s=-\infty}^{\infty} h_{k+s} h_s = \left| \sum_{s=-\infty}^{\infty} h_s \mathrm{e}^{2\pi \mathrm{i}\lambda s} \right|^2. \tag{4.4.20}$$

Заметим далее, что поскольку $\mathrm{e}^{2\pi i \lambda k},\ k = 0, \pm 1, \pm 2, \ldots$ — полная система ортонормальных функций в $L_2(-1/2, 1/2)$, то из (4.4.18) получаем

$$\sum_{s=-\infty}^{\infty} \mathrm{e}^{2\pi i \lambda s} h_s = \sqrt{f_Y(\lambda)}$$

и, следовательно, подставляя это равенство в (4.4.20) получаем $f(\lambda) = f_Y(\lambda)$. ■

Это теорема показывает, что стационарные последовательности можно генерировать пропуская белый шум через инвариантный во времени линейный фильтр. Это фильтр легко вычисляется, но является физически не реализуемым.

Более интересной является задача представления стационарной последовательности как отклика физически реализуемого фильтра. Следующая теорема показывает как такой фильтр в принципе можно было бы построить.

Теорема 4.4.3 *Предположим, что*

$$\int_{-1/2}^{1/2} \log^2[f_Y(\lambda)]\, d\lambda < \infty.$$

Пусть величины $h_k,\ k \geq 0$ *определены из равенства*

$$\sum_{k=0}^{\infty} h_k z^k = \exp\left\{ \sum_{k=0}^{\infty} \phi_k z^k \right\},\ \text{где } \phi_k = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi \mathrm{i}\lambda k} \log[f_Y(\lambda)]\, d\lambda. \tag{4.4.21}$$

Тогда стационарная последовательность

$$Y_k = \sum_{s=-\infty}^{k} h_{k-s} \xi_s,$$

где ξ_s *— стандартный белый гауссовский шум, имеет спектральную плотность* $f_Y(\lambda)$.

Доказательство. В силу (4.4.19) и (4.4.20) достаточно проверить, что

$$f_Y(\lambda) = \left|\sum_{k=0}^{\infty} \mathrm{e}^{2\pi \mathrm{i}\lambda k} h_k\right|^2$$

при h_k из (4.4.21).

Заметим, что поскольку $\phi_k = \phi_{-k}$, то в силу полноты системы функций $\exp(2\pi \mathrm{i} k\lambda), k = 0, \pm 1, \ldots$ в $L_2(-1/2, 1/2)$ имеем

$$\log f_Y(\lambda) = \sum_{k=-\infty}^{\infty} \phi_k \exp(2\pi \mathrm{i} k\lambda) = \phi_0 + \sum_{k=1}^{\infty} \phi_k \bigl[\exp(2\pi \mathrm{i} k\lambda) + \exp(-2\pi \mathrm{i} k\lambda)\bigr].$$

Поэтому из (4.4.21) получим, что

$$\begin{aligned}\left|\sum_{k=0}^{\infty} \mathrm{e}^{2\pi \mathrm{i}\lambda k} h_k\right|^2 &= \left|\exp\left\{\sum_{k=0}^{\infty} \phi_k \mathrm{e}^{2\pi \mathrm{i}\lambda k}\right\}\right|^2 \\ &= \exp\left\{\phi_0 + \sum_{k=1}^{\infty} \phi_k \bigl[\exp(2\pi \mathrm{i} k\lambda) + \exp(-2\pi \mathrm{i} k\lambda)\bigr]\right\} \\ &= \exp\bigl\{\log[f_Y(\lambda)]\bigr\} = f_Y(\lambda). \quad \blacksquare\end{aligned}$$

Еще один способ генерирования стационарных гауссовских последовательностей основан на так называемом спектральном представлении. Для этого представления нам нужно прежде всего определить белый гауссовский шум с непрерывным временем. Рассмотрим последовательность случайных процессов с непрерывным временем

$$\xi_n(t) = \sqrt{n} \sum_{s=-\infty}^{\infty} \xi_s \mathbf{1}\left\{\frac{s}{n} < t \le \frac{s+1}{n}\right\}$$

Очевидно, что при фиксированном n процесс $\xi_n(t)$ это случайный процесс в обычном смысле. Давайте посмотрим в каком смысле можно было бы определить случайный процесс $\xi(t) = \lim\limits_{n\to\infty} \xi_n(t)$. Ясно, что при любом фиксированном t этот предел не существует.

Пусть $\phi \in \mathcal{D}$, где $\mathcal{D}$ — пространство бесконечно дифференцируемых функций с финитным носителем. Рассмотрим последовательность линейных функционалов

$$\begin{aligned}\zeta_n = \langle \phi, \xi_n \rangle = \int_{-\infty}^{\infty} \phi(t)\xi_n(t)\, dt &= \sqrt{n} \sum_{s=-\infty}^{\infty} \xi_s \int_{s/n}^{(s+1)/n} \phi(u)\, du \\ &= \frac{1}{\sqrt{n}} \sum_{s=-\infty}^{\infty} \xi_s \left(n \int_{s/n}^{(s+1)/n} \phi(u)\, du\right).\end{aligned}$$

Понятно, что ζ_n — последовательность гауссовских случайных величин с нулевым средним и дисперсией

$$\begin{aligned}\mathbf{E}\langle\phi,\xi_n\rangle^2 =& \frac{1}{n}\sum_{s=-\infty}^{\infty}\left[n\int_{s/n}^{(s+1)/n}\phi(u)\,du\right]^2\\ =& \frac{1}{n}\sum_{s=-\infty}^{\infty}\left[\phi\left(\frac{s}{n}\right)+\frac{1}{n}O\left(\phi'\left(\frac{s}{n}\right)\right)\right]^2\\ =& \int_{-\infty}^{\infty}\phi^2(u)\,du+o(1),\ n\to\infty.\end{aligned}$$

Поэтому существует для любой функции $\phi\in\mathcal{D}$ существует $\lim_{k\to\infty}\langle\phi,\xi_k(\cdot)\rangle$, который является гауссовской случайной величиной с нулевым средним и дисперсией $\|\phi\|^2$. Отсюда ясно также, что этот предел существует для любой функции $\phi\in L_2(-\infty,\infty)$.

Действуя аналогичным образом, нетрудно также проверить, что при $n\to\infty$

$$\lim_{n\to\infty}\mathbf{E}\langle\phi_1,\xi_n\rangle\langle\phi_2,\xi_n\rangle=\langle\phi_1,\phi_2\rangle.$$

Таким образом, при $n\to\infty$ любая последовательность конечномерных векторов $\{\langle\phi_1,\xi_n\rangle,\dots,\langle\phi_N,\xi_n\rangle\}$ сходится к гауссовскому вектору $\{\zeta_1,\dots,\zeta_N\}$, имеющему ковариационную матрицу $B_{ij}=\langle\phi_i,\phi_j\rangle$.

Это свойство лежит в основе определения белого гауссовского шума с непрерывным временем.

Определение 7 *Обобщенный случайный процесс $\xi(\cdot)$ называется белым гауссовским шумом, если для любых функций $\phi_i\in L_2(-\infty,\infty)$, $i=1\dots,N$, случайный вектор $\big(\langle\phi_1,\xi\rangle,\dots,\langle\phi_N,\xi\rangle\big)^\top$ является гауссовским с нулевым средним и ковариационной матрицей $B_{ij}=\langle\phi_i,\phi_j\rangle$.*

Формально это определение означает, что

$$\mathbf{E}\xi(t)\xi(s)=\delta(t-s).$$

Действительно, действуя формально, находим

$$\begin{aligned}\mathbf{E}\langle\phi,\xi\rangle\langle\phi,\xi\rangle =& \int\int\phi(u)\psi(v)\mathbf{E}\xi(u)\xi(v)\,du\,dv\\ &\int\int\phi(u)\psi(v)\delta(u-v)\,du\,dv=\langle\phi,\psi\rangle.\end{aligned}$$

Поэтому очевидно, что спектральная плотность такого процесса равна

$$f(\lambda)=\int_{-\infty}^{\infty}\mathrm{e}^{2\pi\mathrm{i}t\lambda}\mathbf{E}\xi(t)\xi(0)\,dt=\int_{-\infty}^{\infty}\mathrm{e}^{2\pi\mathrm{i}t\lambda}\delta(t)\,dt=1.$$

С помощью белого гауссовского шума $\xi(\cdot)$ можно легко строить стационарные гауссовскими последовательности с заданной спектральной плотностью.

Теорема 4.4.4 *Пусть Y_k — стационарная последовательность со спектральной плотностью $f_Y(\lambda)$. Тогда*

$$\begin{aligned} Y_k = &\int_{-1/2}^{1/2} \cos[2\pi k\lambda]\sqrt{f_Y(\lambda)}\xi(\lambda)\,d\lambda \\ &+ \int_{-1/2}^{1/2} \sin[2\pi k\lambda]\sqrt{f_Y(\lambda)}\xi'(\lambda)\,d\lambda, \end{aligned} \tag{4.4.22}$$

где $\xi(t)$ и $\xi'(t)$ — независимые белые гауссовские шумы.

Доказательство. Действительно, в силу определения белого гауссовского шума, находим

$$\begin{aligned} \mathbf{E}Y_kY_s &= \int_{-1/2}^{1/2} \big[\cos(2\pi\lambda k)\cos(2\pi\lambda s) + \sin(2\pi\lambda k)\sin(2\pi\lambda s)\big] f_Y(\lambda)\,d\lambda \\ &= \int_{-1/2}^{1/2} \cos[2\pi\lambda(k-s)] f_Y(\lambda)\,d\lambda = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi\mathrm{i}\lambda(k-s)} f_Y(\lambda)\,d\lambda. \quad \blacksquare \end{aligned}$$

Формула (4.4.22) называется спектральным представлением стационарной гауссовской последовательности.

4.4.2 Сглаживание

Предположим что наблюдается гауссовская стационарная последовательность Y_k, $k = 0, \pm1, \pm2, \ldots$, имеющая следующую структуру:

$$Y_k = S_k + N_k,$$

где

- S_k — полезный сигнал, представляющий собой стационарную гауссовскую последовательность с нулевым средним и известной спектральной плотностью $f_S(\lambda)$,
- N_k — шум, представляющий собой стационарную гауссовскую последовательность, независящую от последовательности S и имеющую нулевое среднее и известную спектральную плотность $f_N(\lambda)$. Довольно естественно считать, что полезный сигнал и шум являются независимыми последовательностями.

Задача состоит в том, чтобы по наблюдениям $Y_k,\ 0, \pm 1, \pm 2, \ldots$ оценить значение сигнала S_m. При этом мы измеряем качество оценки $\bar{S}(Y)$ среднеквадратичным риском

$$r(\bar{S}) = \mathbf{E}[\bar{S}(Y) - S_m]^2$$

и наша цель найти наилучшую оценку

$$\hat{S}_m(Y) = \arg\min_{\bar{S}} r(\bar{S}).$$

Теорема 4.4.5 *Байесовская оценка S_m имеет следующий вид:*

$$\hat{S}_m(Y) = \sum_{k=-\infty}^{\infty} h_{m-k} Y_k, \tag{4.4.23}$$

где

$$h_j = \int_{-1/2}^{1/2} \frac{f_S(\lambda)}{f_S(\lambda) + f_N(\lambda)} \mathrm{e}^{-2\pi \mathrm{i} j \lambda}\, d\lambda,$$

а ее байесовский риск вычисляется как

$$r(\hat{S}_m) = \mathbf{E}[S_m - \hat{S}_m]^2 = \int_{-1/2}^{1/2} \frac{f_S(\lambda) f_N(\lambda)}{f_S(\lambda) + f_N(\lambda)}\, d\lambda. \tag{4.4.24}$$

Доказательство. Согласно теореме 4.3.4 байесовская оценка является линейной по наблюдениям X

$$\hat{S}_m = \sum_{k=-\infty}^{\infty} h_{m,k} Y_k,$$

где $h_{m,k}$ определяются из уравнения Винера-Хопфа

$$\mathbf{E}\Big[S_m - \sum_{i=-\infty}^{\infty} h_{m,k} Y_k\Big] Y_j = 0, \quad j = 0, \pm 1, \ldots.$$

Используя свойство стационарности последовательностей S_k и N_k и их независимость, перепишем эти уравнение в следующем виде:

$$\mathbf{E} S_0 S_{m-j} - \sum_{k=-\infty}^{\infty} h_{m,k} \big[\mathbf{E} S_0 S_{k-j} + \mathbf{E} N_0 N_{k-j}\big] = 0. \tag{4.4.25}$$

Заметим, что

$$\sum_{j=-\infty}^{\infty} \exp(-2\pi \mathrm{i} j\lambda)\mathbf{E}S_0 S_{m-j} = \sum_{k=-\infty}^{\infty} \exp[2\pi \mathrm{i}(k-m)\lambda]\mathbf{E}S_0 S_k$$
$$= \exp(-2\pi m\lambda) f_S(\lambda).$$

и по тем же соображениям

$$\sum_{j=-\infty}^{\infty} \exp(-2\pi \mathrm{i} j\lambda)\mathbf{E}N_0 N_{m-j} = \exp(-2\pi m\lambda) f_N(\lambda).$$

Таким образом, умножая уравнения (4.4.25) на $\exp(-2\pi \mathrm{i} j\lambda)$ и суммируя их по j приходим к

$$\mathrm{e}^{-2\pi \mathrm{i} m\lambda} f_S(\lambda) = [f_S(\lambda) + f_N(\lambda)] \sum_{k=-\infty}^{\infty} \mathrm{e}^{-2\pi \mathrm{i} k\lambda} h_{m,k}$$

Отсюда очевидно получим

$$\sum_{k=-\infty}^{\infty} \mathrm{e}^{-2\pi \mathrm{i}(k-m)\lambda} h_{m,k} = \frac{f_S(\lambda)}{f_S(\lambda) + f_N(\lambda)}.$$

Поскольку $\exp(-2\pi \mathrm{i} k\lambda)$, $k = 0, \pm 1, \ldots$ — полная система ортонормальных функций в $L_2(-1/2, 1/2)$, то из этого соотношения получаем (4.4.23).

Для доказательства (4.4.24) воспользуемся еще раз теоремой 4.3.4. Из нее и из тождества Парсеваля находим

$$\begin{aligned} r(\hat{S}_m) =& \mathbf{E}S_m^2 - \sum_{k=-\infty}^{\infty} h_{m-k}\mathbf{E}Y_k S_m \\ =& \mathbf{E}S_m^2 - \sum_{k=-\infty}^{\infty} h_{m-k}\mathbf{E}S_0 S_{m-k} = \mathbf{E}S_m^2 - \sum_{k=-\infty}^{\infty} h_k \mathbf{E}S_0 S_k \\ =& \int_{-1/2}^{1/2} f_S(\lambda)\, d\lambda - \int_{-1/2}^{1/2} \frac{f_S^2(\lambda)}{f_N(\lambda) + f_S(\lambda)}\, d\lambda \\ =& \int_{-1/2}^{1/2} \frac{f_S(\lambda) f_N(\lambda)}{f_N(\lambda) + f_S(\lambda)}\, d\lambda. \quad \blacksquare \end{aligned}$$

4.4.3 Интерполяция

Задача интерполяции состоит в том, чтобы по наблюдениям случайной гауссовской стационарной последовательности

$$\dots, Y_{-2}, Y_{-1}, \ , Y_1, Y_2, \dots$$

оценить величину Y_0. При этом предполагается, что спектральная плотность последовательности $f_Y(\lambda)$ известна.

Теорема 4.4.6 *Предположим, что $f_Y(\lambda) \geq \delta > 0$ при всех λ. Тогда байесовская оценка Y_0 имеет вид*

$$\hat{Y}_0 = \sum_{k \neq 0} h_k Y_k,$$

где

$$h_k = \int_{-1/2}^{1/2} \mathrm{e}^{-2\pi \mathrm{i} k\lambda} \left[1 - \frac{\alpha}{f_X(\lambda)}\right] d\lambda, \qquad \alpha = \left[\int_{-1/2}^{1/2} \frac{1}{f_X(\lambda)}\, d\lambda\right]^{-1}.$$

При этом

$$\mathbf{E}[\hat{Y}_0 - Y_0]^2 = \alpha.$$

Доказательство. Оптимальные величины h_k должны удовлетворять уравнению Винера-Хопфа

$$\mathbf{E}\left(Y_0 - \sum_{k \neq 0} h_k Y_k\right) Y_s = 0, \ s = \pm 1, \pm 2, \dots$$

или, что эквивалентно

$$R(s) - \sum_{k \neq 0} h_k R(s-k) = 0, \ s = \pm 1, \pm 2, \dots,$$

где $R(l) = \mathbf{E} Y_0 Y_l$. Считая, что $h_0 = 0$, перепишем это уравнение в следующем виде

$$R(s) = \sum_{k=-\infty}^{\infty} h_k R(k-s), \ s = \pm 1, \pm 2, \dots$$

Отсюда сразу же получим,

$$\sum_{s \neq 0} R(s) \exp[2\pi \mathrm{i} s\lambda] = \sum_{k=-\infty}^{\infty} h_k \sum_{s \neq 0} R(k-s) \exp[2\pi \mathrm{i} s\lambda]$$

и, следовательно,

$$f_Y(\lambda) - R(0) = \sum_{k=-\infty}^{\infty} h_k \Bigg[\sum_{s=-\infty}^{\infty} R(k-s) \exp[2\pi \mathrm{i} s] - R(k) \Bigg].$$

Отсюда

$$f_Y(\lambda) - \Bigg[R(0) - \sum_{k=-\infty}^{\infty} h_k R(k) \Bigg] = h(\lambda) f_X(\lambda),$$

где

$$h(\lambda) = \sum_{k \neq 0} h_k \exp[2\pi \mathrm{i} \lambda k].$$

Поэтому обозначив для краткости

$$C = \Bigg[R(0) - \sum_{k=-\infty}^{\infty} h_k R(k) \Bigg],$$

найдем

$$h(\lambda) = 1 - \frac{C}{f_X(\lambda)}.$$

Постоянную C находим из условия

$$h_0 = \int_{-1/2}^{1/2} h(\lambda)\, d\lambda = 0.$$

Следовательно, $C = \alpha$.

Для ошибки интерполяции мы имеем

$$\begin{aligned} \mathbf{E}[\hat{Y}_0 - Y_0]^2 =& \mathbf{E} Y_0^2 - \sum_{k \neq 0} h_k R(k) \\ =& \mathbf{E} Y_0^2 - \int_{-1/2}^{1/2} h(\lambda) f_Y(\lambda)\, d\lambda = \alpha. \quad \blacksquare \end{aligned}$$

4.4.4 Экстраполяция

Предположим, что у нас имеется реализация стационарной гауссовской последовательности последовательности

$$\ldots, Y_{-2}, Y_{-1}, Y_0$$

с известной спектральной плотностью $f_Y(\lambda)$. Задача состоит в том, чтобы по этим данным предсказать значение Y_k для заданного $k \geq 1$.

Предположим, что спектральную плотность можно представить в виде

$$f_Y(\lambda) = \left|\Phi(\mathrm{e}^{2\pi \mathrm{i}\lambda})\right|^2 = \left|\sum_{k=0}^{\infty} \mathrm{e}^{2\pi \mathrm{i}\lambda k}\phi_k\right|^2, \tag{4.4.26}$$

где $\|\phi\| < \infty$ (см. теорему 4.4.3). Как мы видели раньше, это означает, что

$$Y_s = \sum_{l=-\infty}^{s} \phi_{s-l}\xi_l,$$

где ξ_l — стандартный белый гауссовский шум.

Теорема 4.4.7 *Байесовская оценка величины Y_k, построенная по наблюдениям $Y_0, Y_{-1}, Y_{-2}, \ldots$ имеет вид*

$$\hat{Y}_k = \sum_{k=0}^{\infty} h_l Y_{-l},$$

где

$$\hat{h}(\lambda) = \sum_{l=0}^{\infty} \mathrm{e}^{2\pi \mathrm{i}\lambda l} h_l = \mathrm{e}^{2\pi \mathrm{i}\lambda k} \sum_{s=k}^{\infty} \mathrm{e}^{2\pi \mathrm{i}\lambda s}\phi_s \Big/ \sum_{s=0}^{\infty} \mathrm{e}^{2\pi \mathrm{i}\lambda s}\phi_s. \tag{4.4.27}$$

и кроме того

$$\mathbf{E}(Y_k - \hat{Y}_k)^2 = \sum_{s=0}^{k-1} |\phi_s|^2.$$

Доказательство. Из уравнения Винера-Хопфа находим уравнения для $h_l(k),\ l = 0, 1, \ldots$

$$\mathbf{E}\left[Y_k - \sum_{l=0}^{\infty} h_l Y_{-l}\right] Y^*_{-s} = 0,\ s = 0, 1, \ldots.$$

Переходя к спектральной плотности, отсюда находим, что функция $\hat{h}(\lambda)$ должна удовлетворять уравнениям

$$\int_{-1/2}^{1/2} f_Y(\lambda)\mathrm{e}^{-2\pi \mathrm{i}\lambda s}\left[\mathrm{e}^{2\pi \mathrm{i}\lambda k} - \hat{h}(\lambda)\right] d\lambda = 0,\ s \geq 0.$$

Подставляя в это уравнение соотношения (4.4.26) и (4.4.27), находим

$$\int_{-1/2}^{1/2} f_Y(\lambda)\mathrm{e}^{-2\pi\mathrm{i}\lambda s}\Big[\mathrm{e}^{2\pi\mathrm{i}\lambda k}-\hat{h}(\lambda)\Big]\,d\lambda$$
$$=\int_{-1/2}^{1/2}\Big[\sum_{p=0}^{\infty}\mathrm{e}^{-2\pi\mathrm{i}\lambda p}\phi_p\Big]\mathrm{e}^{-2\pi\mathrm{i}\lambda(s-k)}\Big[\sum_{l=0}^{\infty}\mathrm{e}^{2\pi\mathrm{i}\lambda l}\phi_l-\sum_{l=k}^{\infty}\mathrm{e}^{2\pi\mathrm{i}\lambda l}\phi_l\Big]\,d\lambda$$
$$=\int_{-1/2}^{1/2}\Big[\sum_{p=0}^{\infty}\mathrm{e}^{-2\pi\mathrm{i}\lambda p}\phi_p\Big]\mathrm{e}^{-2\pi\mathrm{i}\lambda(s-k)}\Big[\sum_{l=0}^{k-1}\mathrm{e}^{2\pi\mathrm{i}\lambda l}\phi_l\Big]\,d\lambda=0.$$

Ошибка экстраполяции вычисляется следующим образом:

$$\mathbf{E}[Y_k-\bar{X}_k]^2=\mathbf{E}[Y_k-\hat{Y}_k]Y_k$$
$$=\int_{-1/2}^{1/2}\Big[\mathrm{e}^{2\pi\mathrm{i}\lambda k}-\hat{h}(\lambda)\Big]\mathrm{e}^{-2\pi\mathrm{i}\lambda k}f_Y(\lambda)\,d\lambda$$
$$=\int_{-1/2}^{1/2}\Big[1-\mathrm{e}^{-2\pi\mathrm{i}\lambda k}\hat{h}(\lambda)\Big]f_Y(\lambda)\,d\lambda$$
$$=\int_{-1/2}^{1/2}\Big[\sum_{s=0}^{\infty}\mathrm{e}^{2\pi\mathrm{i}\lambda s}\phi_s-\sum_{s=k}^{\infty}\mathrm{e}^{2\pi\mathrm{i}\lambda s}\phi_s\Big]\sum_{s=0}^{\infty}\mathrm{e}^{-2\pi\mathrm{i}\lambda s}\phi_s\,d\lambda$$
$$=\int_{-1/2}^{1/2}\Big[\sum_{s=0}^{k-1}\mathrm{e}^{2\pi\mathrm{i}\lambda s}\phi_s\Big]\sum_{s=0}^{\infty}\mathrm{e}^{-2\pi\mathrm{i}\lambda s}\phi_s\,d\lambda=\sum_{s=0}^{k-1}\phi_s^2.\quad\blacksquare$$

4.4.5 Оценивание параметров спектральной плотности

Во всех предыдущих задачах предполагалось, что спектральная плотность стационарной последовательности известна точно. Это предположение безусловно далеко от реальности. На практике спектральная плотность бывает известна только с точностью до некоторых параметров. Поэтому оценивание этих параметров представляет собой принципиально важную задачу.

Например, во многих практических задачах случайные последовательности Y_s принято описывать так называемыми *авто-регрессионными моделями*, которые имеют следующий вид

$$\sum_{s=0}^{Q}q_sY_{k-s}=\sum_{s=0}^{P}p_s\xi_{k-s},$$

здесь ξ_k — стандартный белый гауссовский шум, а $q_k,\ k=1,\dots Q$ и $p_k,\ k=1,\dots,P$ — некоторые числа. Одним из преимуществ данной

модели является то, что Y_k можно вычислять рекуррентно в реально времени поскольку

$$Y_k = \frac{1}{q_0}\sum_{s=0}^{P} p_s \xi_{k-s} - \frac{1}{q_0}\sum_{s=1}^{Q} q_s Y_{k-s}.$$

Спектральная плотность такой последовательности имеет следующий вид:

$$f_Y(\lambda) = \frac{\left|\sum\limits_{s=0}^{P} \mathrm{e}^{2\pi \mathrm{i}\lambda s} p_s\right|^2}{\left|\sum\limits_{s=0}^{Q} \mathrm{e}^{2\pi \mathrm{i}\lambda s} q_s\right|^2}$$

Такие спектральные плотности образуют очень широкий класс, позволяет хорошо хорошо аппроксимировать случайные последовательности, встречающиеся в прикладных задачах. При этом, как правило, оказывается, что параметры $p_k,\ q_k$ неизвестны и их требуется оценивать по наблюдениям. В качестве наблюдений обычно выступает конечная последовательность $Y^n = \{Y_1, \ldots, Y_n\}$. Задача оценивания параметров $p_k,\ q_k$ часто называется идентификацией модели. При это возникают две кардинально различные задачи

- *при заданных величинах P и Q оценить $p_k,\ k = 1, \ldots, P$ и $q_k,\ k = 1, \ldots, Q$;*
- *оценить размерности модели P и Q.*

Вторая задача существенно выходит за рамки нашего курса и мы ей заниматься не будем. Что касается первой, ее, в принципе, можно было бы решить с помощью метода максимального правдоподобия. При этом можно рассматривать достаточно общую задачу оценивания многомерного параметра $\theta \in \mathbb{R}^d$ спектральной плотности $f_Y(\lambda;\theta)$ по наблюдениям Y^n. Для того, чтобы построить точную оценку максимального правдоподобия нам нужно вычислить плотность распределения наблюдений Y^n. Эту задачу мы уже рассматривали, когда говорили о спектральном представлении гауссовских векторов (см. (4.3.9)). Логарифм отношения правдоподобия вычисляется следующим образом:

$$\log[p(Y^n;\theta)] = -\frac{1}{2}\sum_{i=1}^{n} \frac{\langle Y^n, \phi_i(\theta)\rangle^2}{\lambda_i(\theta)} - \frac{n}{2}\sum_{i=1}^{n} \log[2\pi\lambda_i(\theta)],$$

где $\varphi_i(\theta)$ и $\lambda_i(\theta)$ — собственные векторы и собственные числа корреляционной матрицы $B(\theta)$

$$B_{l,k}(\theta) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi \mathrm{i}(l-k)} f(\lambda;\theta)\, d\lambda,\ l,k = 1,\dots,n.$$

Таким образом, оценка максимального правдоподобия имеет вид

$$\bar{\theta} = \arg\max_{\tilde{\theta}} \log[p(Y^n;\tilde{\theta})].$$

С вычислительной точки зрения эта задача, к сожалению, может оказаться очень тяжелой, поскольку для ее решения придется много раз вычислять собственные числа и собственные векторы матрицы $B(\theta)$.

Естественно хотелось бы избежать подобного рода вычислений. Это можно сделать, если воспользоваться спектральным представлением последовательности Y_k (см. теорему 4.4.4), которое имеет следующий вид:

$$\begin{aligned} Y_k = &\int_{-1/2}^{1/2} \cos[2\pi k\lambda]\sqrt{f(\lambda)}\xi(\lambda)\, d\lambda \\ &+ \int_{-1/2}^{1/2} \cos[2\pi k\lambda]\sqrt{f(\lambda)}\xi'(\lambda)\, d\lambda, \end{aligned} \tag{4.4.28}$$

где $\xi(t)$ и $\xi'(t)$ — независимые стандартные белые гауссовские шумы. Основная идея — это аппроксимировать интегралы в этой формуле конечными суммами. Для этого разобьем отрезок $[-1/2, 1/2]$ на n частей и обозначим для краткости

$$\lambda_s = \frac{s}{n}, \quad s = 0, \pm 1, \dots \pm \frac{n}{2}.$$

Тогда получим

$$\begin{aligned}
&\int_{-1/2}^{1/2} \cos[2\pi k\lambda]\sqrt{f(\lambda)}\xi(\lambda)\,d\lambda \\
&\quad= \sum_{s=-n/2+1}^{n/2} \int_{\lambda_{s-1}}^{\lambda_s} \cos[2\pi k\lambda]\sqrt{f(\lambda)}\xi(\lambda)\,d\lambda \\
&\quad\approx \sum_{s=-n/2+1}^{n/2} \cos[2\pi k\lambda_s]\sqrt{f(\lambda_s)} \int_{\lambda_{s-1}}^{\lambda_s} \xi(\lambda)\,d\lambda \\
&\quad= \frac{1}{\sqrt{n}} \sum_{k=-n/2+1}^{n/2} \cos[2\pi k\lambda_k]\sqrt{f(\lambda_k)}\sqrt{n} \int_{\lambda_{k-1}}^{\lambda_k} \xi(\lambda)\,d\lambda \\
&\quad= \frac{1}{\sqrt{n}} \sum_{s=-n/2+1}^{n/2} \cos[2\pi k\lambda_s]\sqrt{f(\lambda_s)}\xi_{1,s},
\end{aligned} \tag{4.4.29}$$

где

$$\xi_{1,s} = \sqrt{n} \int_{\lambda_{s-1}}^{\lambda_s} \xi(\lambda)\,d\lambda.$$

Аналогичная аппроксимация справедлива и для

$$\begin{aligned}
&\int_{-1/2}^{1/2} \sin[2\pi k\lambda]\sqrt{f(\lambda)}\xi'(\lambda)\,d\lambda \\
&\qquad= \frac{1}{\sqrt{n}} \sum_{s=-n/2+1}^{n/2} \sin[2\pi k\lambda_s]\sqrt{f(\lambda_s)}\xi_{2,s},
\end{aligned} \tag{4.4.30}$$

где

$$\xi_{2,s} = \sqrt{n} \int_{\lambda_{s-1}}^{\lambda_s} \xi'(\lambda)\,d\lambda.$$

Поэтому, объединяя (4.4.29) и (4.4.30), приходим к следующей аппроксимации

$$Y_k \approx \tilde{Y}_k,\ k = 1, \ldots, n, \tag{4.4.31}$$

где

$$\begin{aligned}
\tilde{Y}_k =& \frac{1}{\sqrt{n}} \sum_{s=-n/2+1}^{n/2} \cos[2\pi k\lambda_s]\sqrt{f(\lambda_s)}\xi_{1,s} \\
&+ \frac{1}{\sqrt{n}} \sum_{s=-n/2+1}^{n/2} \sin[2\pi k\lambda_s]\sqrt{f(\lambda_s)}\xi_{2,s},\ k = 1, \ldots, n.
\end{aligned}$$

Плотность распределения $\tilde{p}(x),\ x \in \mathbb{R}^n$ гауссовских случайных величин $\tilde{Y}^n = \{\tilde{Y}_1, \dots, \tilde{Y}_n\}$ легко вычислить.

Теорема 4.4.8

$$\log\bigl[\tilde{p}(y)\bigr] = -\frac{2}{n}\sum_{s=-n/2+1}^{n/2}\frac{1}{f(\lambda_s)}\left|\sum_{k=1}^{n}\exp\bigl[2\pi k\lambda_s\bigr]y_k\right|^2 - \sum_{s=-n/2+1}^{n/2}\log\left[\frac{\pi f(\lambda_s)}{2}\right]. \tag{4.4.32}$$

Доказательство. Заметим, что

- гауссовские случайные величины $\xi_{1,s}$ и $\xi_{2,s},\ s = -n/2+1, \dots, n/2$ независимы и имеют стандартное гауссовское распределение;
- векторы $\psi^{(s)}$ и $\psi^{*(s)}$ с координатами

$$\psi_k^{(s)} = \sqrt{\frac{1}{n}}\cos[2\pi k\lambda_s], \quad \psi_k^{*(s)} = \sqrt{\frac{1}{n}}\sin[2\pi k\lambda_s]$$

ортогональны в $\mathbb{R}^n$ и $\|\psi^{(s)}\|^2 = \|\psi^{*(s)}\|^2 = 1/2$.

Следовательно,

$$\hat{Y}^{(s)} = \sum_{k=1}^{n}\tilde{Y}_k\psi_k^{(s)} = \frac{\sqrt{f(\lambda_s)}}{2}\xi_{1,s}, \quad \hat{Y}^{*(s)} = \sum_{k=1}^{n}\tilde{Y}_k\psi_k^{*(s)} = \frac{\sqrt{f(\lambda_s)}}{2}\xi_{2,s}.$$

Поэтому из независимости $\xi_{1,s},\ \xi_{2,s}$ получаем, что величины $\hat{Y}^{(s)},\ \hat{Y}^{*(s)}$ независимы и имеют плотность распределения

$$p(y_1, y_2) = \exp\left\{-\frac{2y_1^2 + 2y_2^2}{f(\lambda_s)} - \log\left[\frac{\pi f(\lambda_s)}{2}\right]\right\}.$$

Отсюда непосредственно вытекает (4.4.32). ■

Из этой теоремы вытекает следующий метод оценивания неизвестного параметра $\theta \in \Theta \subset \mathbb{R}^p$ спектральной плотности $f(\lambda;\theta)$ на основе наблюдений Y^n

$$\bar{\theta}(Y^n) = \arg\max_{\theta\in\Theta}\left\{-\frac{1}{n}\sum_{s=-n/2+1}^{n/2}\frac{1}{f(\lambda_s;\theta)}\left|\sum_{k=1}^{n}\exp\bigl[2\pi k\lambda_s\bigr]Y_k\right|^2 - \frac{1}{2}\sum_{s=-n/2+1}^{n/2}\log\left[\frac{\pi f(\lambda_s;\theta)}{2}\right]\right\}.$$

Этот метод называется методом Витла. Заметим, что самая сложная операция при его реализации это вычисление быстрого преобразования Фурье.

4.5 Оптимальность δ - метода

В основе метода максимального правдоподобия, как мы видели, лежит δ - метод. В этом разделе мы докажем оптимальность этого метода.

Пусть Y^n повторная выборка из закона распределения с плотностью с неизвестной плотностью $p(x),\ x \in \mathbb{R}^1$. Наша задача оценить линейный функционал

$$\phi(p) = \int_{\mathbb{R}} \phi(x) p(x)\, dx$$

на основе наблюдений Y^n. Для решения этой задачи воспользуемся δ - методом, т. е. оценим этот функционал как

$$\bar{\phi}(Y^n) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i).$$

Заметим, что

$$\begin{aligned} \mathbf{E}\left[\bar{\phi}(Y^n) - \phi(p)\right]^2 =& \frac{1}{n}\left[\int \phi^2(x)\, p(x)\, dx - \left(\int \phi(x)\, dx\right)^2\right] \\ \stackrel{\text{def}}{=}& \frac{\sigma^2(\phi, p)}{n}. \end{aligned} \tag{4.5.33}$$

В следующей теореме будет показано, что $\bar{\phi}_{Y^n}$ является в некотором смысле асимптотически оптимальной оценкой. Доказать такой факт можно только при условии, что плотность распределения $p_Y(\cdot)$ является неизвестной. Математически это утверждение можно следующим образом. Пусть у нас имеется некоторая известная плотность распределения $p^\circ(\cdot)$. Определим окрестность этой плотности как

$$D(p^\circ, \epsilon) = \left\{ p : \int_{\mathbb{R}} |p(z) - p^\circ(z)|\, dz \le \varepsilon,\ p(x) \ge 0,\ \int_{\mathbb{R}} p(x)\, dx = 1 \right\}.$$

Другими словами $D(p^\circ, \epsilon)$ это — шар радиуса 2ε в метрике полной вариации.

Теорема 4.5.1 *Предположим, что плотность $p^\circ(\cdot)$ такова, что*

$$\int \phi^2(x)p^\circ(x)\,dx < \infty, \quad \sigma^2(p^\circ,\phi) > 0$$

и что последовательность $\varepsilon_n \to 0$ такова, что

$$\lim_{n\to\infty} n\varepsilon_n^2 = \infty.$$

Тогда

$$\lim_{n\to\infty} \inf_{\hat\phi} \sup_{p\in D(p^\circ,\varepsilon_n)} n\mathbf{E}\,[\hat\phi(Y^n) - \phi(p)]^2 \geq \sigma^2(\phi, p^\circ),$$

где inf *вычисляется по всем оценкам функционала $\phi(p)$.*

Доказательство. Основная идея в доказательстве этого результата — вложить в окрестность некоторой параметрическое семейство плотностей. Пусть

$$\phi^A(x) = \phi(x)1\{|\phi(x)| \leq A\}.$$

Рассмотрим следующее параметрическое семейство плотностей

$$p(x;\theta) = C(\theta)p^\circ(x)[1 + \theta\phi^A(x)],$$

где

$$C(\theta) = \left(1 + \theta\int_{\mathbb{R}} p^\circ(u)\phi^A(u)\,du\right)^{-1}, \quad |\theta| \leq \delta_n;$$

здесь последовательность δ_n такова, что выполняются следующие условия:

$$\begin{aligned} &\delta_n \leq \frac{\epsilon_n}{A(2+\epsilon_n)},\\ &\lim_{n\to\infty} n\delta_n^2 = \infty,\\ &\lim_{n\to\infty} n\delta_n^4 = 0. \end{aligned} \tag{4.5.34}$$

Очевидно, что функция $p_\theta(x)$ будет плотностью при всех достаточно больших n. Заметим также, что

$$\begin{aligned} \int_{\mathbb{R}} |p(z;\theta) - p^\circ(z)|\,dz &= \int_{\mathbb{R}} \left|1 - C(\theta)]p^\circ(x) + C(\theta)\theta p^\circ(z)\phi^A(z)\right|\,dz\Big| \\ &\leq |1 - C(\theta)| + A|\theta||C(\theta)| \leq \frac{2A|\theta|}{1 - A|\theta|}. \end{aligned}$$

Таким образом, $p_\theta \in D(p^\circ, \varepsilon_n)$ и, следовательно, мы получаем границу снизу для минимаксного риска

$$\begin{aligned}\rho_n &\stackrel{\text{def}}{=} \inf_{\hat\phi} \sup_{p\in K(p^\circ,\varepsilon_n)} n\mathbf{E}\,[\hat\phi(Y^n) - \phi(p)]^2 \\ &\geq \inf_{\hat\phi} \sup_{|\theta|\leq\delta_n} n\mathbf{E}\left\{\hat\phi(Y^n) - \phi[p(\cdot;\theta)]\right\}^2 \\ &\geq \inf_{\hat\phi} n\int_{\mathbb{R}} \pi_n(\theta)\mathbf{E}\left\{\hat\phi(Y^n) - \phi[p(\cdot;\theta)]\right\}^2 d\theta,\end{aligned} \tag{4.5.35}$$

где $\pi_n(\theta)$ — любая плотность распределения с носителем на отрезке $[-\delta_n, \delta_n]$. Далее будем считать, что

$$\pi_n(\theta) = \frac{1}{2\varepsilon_n}\left[1 + \cos\left(\frac{\pi\theta}{\delta_n}\right)\right]\mathbf{1}\big\{|\theta| \leq \delta_n\big\}. \tag{4.5.36}$$

Посмотрим теперь как зависит $\phi(p_\theta)$ от θ. Имеем

$$\phi[p(\cdot;\theta)] = C(\theta)\int_{\mathbb{R}} p^\circ(x)\phi(x)\,dx + \theta C(\theta)\int_{\mathbb{R}} p^\circ(x)\phi(x)\phi^A(x)\,dx. \tag{4.5.37}$$

Воспользовавшись формулой Тейлора тем, что $|\theta| \leq \delta_n$, получим, что

$$C(\theta) = 1 - \theta\int_{\mathbb{R}} p^\circ(x)\phi^A(x)\,dx + O\big(\delta_n^2\big).$$

Поставляя это соотношение в (4.5.37), получаем

$$\phi[p(\cdot;\theta)] = \phi(p^\circ) + \theta\sigma^2(p^\circ, \phi, \phi^A) + O\big(\delta_n^2\big). \tag{4.5.38}$$

где

$$\begin{aligned}\sigma^2(p^\circ, \phi, \phi^A) = &\int p^\circ(x)\phi(x)\phi^A(x)\,dx \\ &- \int p^\circ(x)\phi(x)\,dx \cdot \int p^\circ(x)\phi^A(x)\,dx.\end{aligned}$$

Заметим также, что любых чисел x, y справедливо следующее неравенство:

$$(x+y)^2 = x^2 + y^2 + 2\sqrt{\gamma}x\cdot\frac{y}{\sqrt{\gamma}} \geq (1-\gamma)x^2 - \gamma^{-1}y^2.$$

Поэтому применяя это неравенство, из и получим что для любого $\gamma > 0$ выполняется неравенство

$$\rho_n \geq (1-\gamma)\sigma^4(p^\circ, \phi, \phi^A)\inf_{\hat\theta} n\int_{\mathbb{R}} \pi_n(\theta)\mathbf{E}(\hat\theta - \theta)^2 - n\gamma^{-1}O(\delta_n^4). \tag{4.5.39}$$

Чтобы продолжить это неравенство, применяем неравенство Ван Триса (4.2.3) и получаем

$$n\int_{\mathbb{R}}\pi_n(\theta)\mathbf{E}(\hat{\theta}-\theta)^2\geq\frac{n}{nI_{p,n}+I_{\pi,n}},\tag{4.5.40}$$

где

$$\begin{aligned}I_{p,n}&=\int_{\mathbb{R}}\int_{\mathbb{R}}\pi_n(\theta)\frac{1}{p_\theta(x)}\left[\frac{d}{d\theta}p_\theta(x)\right]^2dx\,d\theta\\&=\int_{\mathbb{R}}\pi_n(\theta)\int_{\mathbb{R}}p^\circ(x)\frac{[C'(\theta)(1+\theta\phi^A(x))+C(\theta)\phi^A(x)]^2}{C(\theta)[1+\theta\phi^A(x)]}\,dx\,d\theta\end{aligned}\tag{4.5.41}$$

и

$$I_{\pi,n}=\int_{\mathbb{R}}\frac{[\pi_n'(\theta)]^2}{\pi_n(\theta)}\,d\theta=\frac{1}{\delta_n^2}\int_{-1}^{1}\frac{\sin^2(\pi x)}{1+\cos(\pi x)}\,dx=O\left(\frac{1}{\delta_n^2}\right).\tag{4.5.42}$$

Заметим, что равномерно по θ таким, что $|\theta|\leq\delta_n$ выполнены следующие соотношения:

$$\begin{aligned}C(\theta)=&1+O(\delta_n),\\C'(\theta)=&-(1+O(\delta_n))\int_{\mathbb{R}}\phi^A(x)p^\circ(x)\,dx.\end{aligned}$$

Отсюда и из (4.5.41) находим

$$\lim_{n\to\infty}I_{p,n}=\int_{\mathbb{R}}\left[\phi^A(x)-\int_{\mathbb{R}}\phi^A(u)p^\circ(u)\,du\right]^2p^\circ(x)\,dx=\sigma^2(p^\circ,\phi^A).$$

В силу условий (4.5.34) и (4.5.42)

$$\lim_{n\to\infty}\frac{I_{\pi,n}}{n}=0.$$

Поэтому

$$\lim_{n\to\infty}n\int_{\mathbb{R}}\pi_n(\theta)\mathbf{E}(\hat{\theta}-\theta)^2\geq\frac{1}{\sigma^2(p^\circ,\phi^A)}\tag{4.5.43}$$

Далее, поскольку (см. условия (4.5.34))

$$\lim_{n\to\infty}nO(\delta_n^4)=0$$

из (4.5.43) и (4.5.40) получаем

$$\lim_{n\to\infty}\rho_n\geq(1-\gamma)\frac{\sigma^4(p^\circ,\phi,\phi^A)}{\sigma^2(p^\circ,\phi^A)}$$

и, переходя в этом неравенстве к пределу при $A\to\infty$ и $\gamma\to0$, завершаем доказательство теоремы. ■

4.6 Вычисление байесовских оценок в негауссовских моделях

Если размерность оцениваемого вектора велика, то вычисление байесовской оценки может представлять в общем случае достаточно сложную задачу, поскольку оно сводится к вычислению многомерных интегралов. Очень часто для решения этой задачи используется

Алгоритм Метрополиса-Хастингса

Этот метод генерирует случайные величины $X_1, X_2, \ldots$, имеющие плотность

$$q(\theta) \stackrel{\text{def}}{=} \pi(\theta)p(Y^n;\theta),\ \theta \in \mathbb{R}^d.$$

Подчеркнем, что при этом не предполагается, что случайные величины $X_1, X_2, \ldots$ должны быть независимыми. Отказ от требования независимости позволяет генерировать требуемые величины сравнительно просто.

Идея алгоритма Метрополиса-Хастингса состоит в том, чтобы построить марковскую цепь $X_k \in \mathbb{R}^d$, у которой стационарной плотностью была бы плотность $q(\cdot)$. Пусть $P(x,y)$ — переходная плотность этой цепи, то есть

$$\mathbf{P}\{X_n \in [y, y+dy] \big| X_{n-1} = x\} = P(x,y)\,dy.$$

Здесь и далее $\mathbf{P}\{A\big|B\}$ означает условную вероятность выполнения события A при условии, что выполнено событие B.

Обозначим

$$\mathbf{P}\{X_n \in [y, y+dy] \big| X_0 = x\} = P^n(x,y)\,dy$$

и предположим, что существует

$$\lim_{n\to\infty} P^n(x,y) = q(y).$$

Тогда очевидно, что $q(\cdot)$ находится как решение уравнения

$$q(y) = \int_{\mathbb{R}^d} P(x,y)q(x)\,dx$$

поскольку

$$P^{n+1}(x,y) = \int_{\mathbb{R}^d} P^n(x,\xi)P(\xi,y)\,d\xi.$$

Пусть $v(x,y),\ x,y \in \mathbb{R}^d$ — некоторая функция такая, что

1. $v(x,y) > 0$ для всех $x, y \in \mathbb{R}^d$.
2. $\int\limits_{\mathbb{R}^d} v(x,y)\,dy = 1$.
3. Существует простой метод генерирования независимых случайных чисел $Z_k(x)$, распределенных с плотностью $v(x,\cdot)$

$$\frac{d}{dy}\mathbf{P}\big\{Z_k(x) \le y\big\} = v(x,y).$$

Обозначим

$$\alpha_q(x,y) = \min\left\{\frac{q(y)v(y,x)}{q(x)v(x,y)}, 1\right\}.$$

Алгоритм

1. Пусть $X_0 \in \mathbb{R}^k$ -- любой вектор и $i = 0$.
2. Генерируем случайную величину $Z_i(X_i)$ с плотностью $v(X_i,\cdot)$ и независимую равномерно распределенную на $[0,1]$ случайную величину U_i. Вычисляем

$$X_{i+1} = Z_i(X_i)\mathbf{1}\big\{U_i \le \alpha_q(X_i, Z_i)\big\} + X_i\mathbf{1}\big\{U_i > \alpha_q(X_i, Z_i)\big\}.$$

3. Полагаем i=i+1 и переходим к второму пункту.

Теорема 4.6.1 *Стационарная плотность распределения марковской цепи X_k равна $q(x)$, $x \in \mathbb{R}^d$.*

Доказательство. Положим

$$s(x,y) = \begin{cases} v(x,y)\alpha_q(x,y), & x \ne y \\ 0, & x = y. \end{cases}$$

Тогда переходная плотность определяется следующим образом:

$$P_q(x,y) = s(x,y) + \delta(x-y)r(x), \tag{4.6.44}$$

где $r(x)$ — вероятность события что алгоритм останется в точке x

$$r(x) = 1 - \int\limits_{\mathbb{R}^d} s(x,\xi)\,d\xi. \tag{4.6.45}$$

Заметим, что

$$q(x)s(y,x) = q(y)s(x,y) \tag{4.6.46}$$

поскольку

$$q(x)v(x,y)\min\left\{\frac{q(y)v(y,x)}{q(x)v(x,y)},1\right\} = q(y)v(y,x)\min\left\{\frac{q(x)v(x,y)}{q(y)v(y,x)},1\right\}.$$

Из (4.6.45–4.6.46) находим

$$\begin{aligned}\int_{\mathbb{R}^d} P_q(x,y)q(x)\,dx &= \int_{\mathbb{R}^d} s(x,y)q(x)\,dx + q(y)r(y)\\ &= \int_{\mathbb{R}^d} q(y)s(y,x)\,dx + q(y)\left[1 - \int_{\mathbb{R}^d} s(y,x)\,dx\right]\\ &= q(y)\int_{\mathbb{R}^d} s(y,x)\,dx + q(y) - q(y)\int_{\mathbb{R}^d} s(y,x)\,dx = q(y). \quad \blacksquare\end{aligned}$$

Пример. Оценивание параметров смеси распределений

Пусть $Y^n = \{Y_1,\dots,Y_n\}$ — повторная выборка из распределения с плотностью

$$p(x;\theta) = \sum_{i=1}^{d-1}\theta_i p_i^\circ(x) + \left(1 - \sum_{i=1}^{d-1}\theta_i\right)p_d^\circ(x), \quad x\in\mathbb{R}^1.$$

Плотности $p_i^\circ(x)$ предполагаются известными и задача состоит в том, чтобы оценить вектор θ, принадлежащий симплексу

$$\Theta = \left\{\theta : \theta_i \ge 0, \sum_{i=1}^{d-1}\theta_i \le 1\right\}.$$

Предположим, что θ — случайный вектор, равномерно распределенный на Θ. Наша цель — построить байесовскую оценку этого вектора при квадратичной функции потерь.

$$\hat\theta_i(Y^n) = \int_\Theta \theta_i q(\theta;Y^n)\,d\theta_1\cdots d\theta_d,$$

где

$$q(\theta;Y^n) = \prod_{k=1}^n p(Y_k;\theta)\Big/\int_\Theta\prod_{s=1}^n p(Y_s;t)\,dt_1\cdots dt_d.$$

Как мы видели эта оценка минимизирует среднеквадратичный риск, точнее для нее справедливо следующее равенство:

$$\inf_{\tilde\theta} \mathbf{E}\big\|\theta - \tilde\theta(Y^n)\big\|^2 = \mathbf{E}\big\|\theta - \hat\theta(Y^n)\big\|^2 = R.$$

Отметим также, что величину R можно оценить как

$$\hat R(Y^n) = \int_\Theta \big\|\theta - \hat\theta(Y^n)\big\|^2 q(\theta; Y^n) d\theta_1 \cdots d\theta_d.$$

При достаточно больших размерностях d интегрирование на симплексе Θ — трудная задача. Обойти эту неприятность можно если воспользоваться алгоритмом Метрополиса-Хастигса для построения марковской цепи X_k, имеющей стационарную плотность $q(\theta, Y^n)$. Если такую цепь удается построить, то согласно закону больших чисел для марковских цепей

$$\hat\theta(Y^n) = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^m X_i$$
$$\hat R(Y^n) = \lim_{m\to\infty} \left\{ \frac{1}{m}\sum_{i=1}^m \big\|X_i\big\|^2 - \left\|\frac{1}{m}\sum_{i=1}^m X_i\right\|^2 \right\}.$$

Чтобы построить требуемую цепь необходимо подобрать соответствующим образом функцию $q(\cdot,\cdot)$. Идея подбора этой функции основана на асмиптотическом поведении апостериорной плотности $q(\cdot, Y^n)$. Хорошо известно, что при больших n

$$q(\theta; Y^n) \asymp \exp\left\{ -\frac{1}{2}\big\langle I(\hat\theta)[\theta - \hat\theta(Y^n)], [\theta - \hat\theta(Y^n)]\big\rangle \right\},$$

где $I(\theta)$ — фишеровская информационная матрица с элементами

$$\begin{aligned} I_{kl}(\theta) =& n\int_{-\infty}^{\infty} \frac{\partial p(\theta;x)}{\partial\theta_k}\frac{\partial p(\theta;x)}{\partial\theta_l} p^{-1}(\theta;x)\,dx \\ =& n\int_{-\infty}^{\infty} \frac{[p_k(x) - p_d(x)][p_l(x) - p_d(x)]}{p(x;\theta)}\,dx, \quad k,l = 1,\ldots,d-1. \end{aligned}$$

Поэтому довольно естественно взять в качестве функции $v(\cdot,\cdot)$ в алгоритме Метрополиса-Хастингса

$$v(x,y) = \exp\left\{ -\frac{1}{2}\big\langle I(x)(y-x), (y-x)\big\rangle \right\}.$$

Это значит, что векторы $Z_k \in R^{d-1}$ в алгоритме являются гауссовскими со средним x и ковариационной матрицей $I^{-1}(x)$.

Глава 5

Статистические тесты

Предположим, что наблюдается случайный вектор $Y^n \in \mathbb{R}^n$, плотность распределения которого $p(\cdot)$ нам неизвестна, но тем не менее нам известно, что

$$p(x) = p(x;\theta), \text{ для некоторго } \theta \in \Theta \subset \mathbb{R}^d;$$

здесь $p(\cdot;\theta),\ \theta \in \Theta$ — семейство известных плотностей распределения.

В теории статистических тестов предполагается, что множество возможных параметров Θ разбито на пересекающиеся компоненты

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \varnothing.$$

Имея это разбиение, наша задача состоит в том, чтобы на основании наблюдений Y^n принять решение о том, какому из множеств Θ_0 или Θ_1 принадлежит неизвестный параметр θ.

Подмножества Θ_0 и Θ_1 называются гипотезами. Точнее, Θ_0 называют нулевой гипотезой, а Θ_1 альтернативой и статистический тест состоит в том, чтобы по наблюдениям Y^n проверить гипотезу

$$H_0 :\ \theta \in \Theta_0 \textit{ против альтернативы } H_1 :\ \theta \in \Theta_1.$$

Гипотеза или альтернатива называется простой если соотвествующее множество Θ_0 или Θ_1 состоит из одной точки. В остальных случаях она называется сложной.

Статистическим тестом называется функция

$$\varphi(x) :\ \mathbb{R}^n \to [0,1],$$

которая имеет следующий смысл:

- если $\varphi(Y^n) = 0$, то принимается гипотеза H_0;
- если $\varphi(Y^n) = 1$, то H_0 отвергается, то есть принимается альтернатива;
- если $\varphi(Y^n) = p$ $(0 < p < 1)$, то альтернатива принимается с вероятностью p. Это означает, что генерируется независимая от выборки случайная величина κ, принимающая значения $\{0, 1\}$ с вероятностями $\{1 - p, p\}$ и принимается гипотезу H_0, если $\kappa = 0$, а альтернативу, если $\kappa = 1$.

При этом можно совершить следующие ошибки:

- отвергнуть H_0, когда H_0 справедлива (*ошибка первого рода*);
- принять H_0, когда H_1 справедлива (*ошибка второго рода*).

Определение 8 *Для заданного статистического теста $\varphi(\cdot)$ его ошибкой первого рода называется величина*

$$\alpha(\varphi, \theta) = \mathbf{E}_\theta \varphi(Y^n), \quad \theta \in \Theta_0,$$

а ошибкой второго рода

$$\beta(\varphi, \theta) = 1 - \mathbf{E}_\theta \varphi(Y^n), \quad \theta \in \Theta_1;$$

здесь

$$\mathbf{E}_\theta \varphi(Y^n) = \int_{\mathbb{R}^n} \varphi(y) p(y; \theta)\, dy.$$

В принципе, цель теории статистических тестов найти функцию $\varphi(\cdot)$, которая минимизировала бы одновременно и $\alpha(\varphi, \theta)$ и $\beta(\varphi, \theta)$. Как правило, оптимальный тест ищется в классе тестов, имеющих заданную вероятность ошибки первого рода α

$$\alpha(\phi, \theta) \leq \alpha,$$

где α — некоторое заданное число, обычно $\alpha = 0,05$. Наилучший тест можно было определить как тест, который с одной стороны удовлетворяет этому неравенству, а с другой стороны минимизирует вероятность ошибки второго рода $\beta(\varphi, \theta)$ при любом $\theta \in \Theta_1$. Эта задача имеет решение только в случае *двух простых гипотез*.

В общем случае у нее нет решения поскольку для ее решения необходимо сравнивать функции (вероятности ошибок второго рода), что невозможно. Поэтому корректно определить оптимальный тест можно только если вместо вероятностей ошибок первого и второго рода ввести некоторый функционал от них.

5.1 Байесовские тесты

Как и при оценивании параметра, мы можем рассмотреть линейный функционал от ошибок первого и второго рода. Иными словами измерять качество теста $\varphi(\cdot)$ следующей величиной:

$$r_\pi[\varphi] = \int_{\Theta_0} \pi(\theta)\mathbf{E}_\theta\varphi(Y^n)\,d\theta + \int_{\Theta_1} \pi(\theta)\mathbf{E}_\theta[1-\varphi(Y^n)]\,d\theta, \tag{5.1.1}$$

где $\pi(\theta)$ — неотрицательная функция такая, что

$$\int_{\mathbb{R}^d} \pi(\theta)\,d\theta = 1.$$

Если используется такой критерий качества, то можно сказать, что мы считаем неизвестный параметр $\theta \in \mathbb{R}^d$ случайной величиной с плотностью $\pi(\theta)$.

Байсовским тестом называется функция

$$\bar\varphi_\pi = \arg\min_\varphi r_\pi[\varphi];$$

здесь минимум вычисляется по всем функциям $\varphi : \mathbb{R}^n \to [0,1]$.

Теорема 5.1.1 *Байесовский тест и его ошибка вычисляются следующим образом:*

$$\bar\varphi_\pi(Y^n) = \mathbf{1}\biggl\{\int_{\Theta_1} \pi(\theta)p(Y^n;\theta)\,d\theta \ge \int_{\Theta_0} \pi(\theta)p(Y^n;\theta)\,d\theta\biggr\}, \tag{5.1.2}$$

$$r_\pi[\bar\varphi_\pi] = \int_{\mathbb{R}^n} \min\biggl\{\int_{\Theta_0} \pi(\theta)p(x;\theta)\,d\theta, \int_{\Theta_1} \pi(\theta)p(x;\theta)\,d\theta\biggr\}\,dx. \tag{5.1.3}$$

Доказательство. Оно аналогично доказательству теоремы 4.1.1 о байесовских оценках. Заметим, что меняя порядок интегрирования в (5.1.1), получим, что для любого теста $\varphi(Y^n)$ справедливо неравенство

$$\begin{aligned} r_\pi[\varphi] &= \int_{\mathbb{R}^n} \biggl\{\varphi(x)\int_{\Theta_0} \pi(\theta)p(x;\theta)\,d\theta + [1-\varphi(x)]\int_{\Theta_1} \pi(\theta)p(x;\theta)\,d\theta\biggr\}\,dx \\ &\ge \int_{\mathbb{R}^n} \min_{y\in[0,1]}\biggl\{y\int_{\Theta_0} \pi(\theta)p(x;\theta)\,d\theta + [1-y]\int_{\Theta_1} \pi(\theta)p(x;\theta)\,d\theta\biggr\}\,dx \\ &= \int_{\mathbb{R}^n} \min\biggl\{\int_{\Theta_0} \pi(\theta)p(x;\theta)\,d\theta, \int_{\Theta_1} \pi(\theta)p(x;\theta)\,d\theta\biggr\}\,dx \\ &= r_\pi[\bar\varphi_\pi]. \quad \blacksquare \end{aligned}$$

5.2 Тесты максимального правдоподобия

Зависимость оптимального статистического теста от произвольной функции $\pi(\theta)$ довольно неприятный факт с практической точки зрения. Поэтому во многих случаях используются тесты максимального правдоподобия. Их идея состоит в замене интегралов, входящих в оптимальный тест (5.1.2) на аппроксимации

$$\int_{\Theta_0} \pi(\theta) p(Y^n;\theta)\, d\theta \approx C_0 \max_{\theta\in\Theta_1} p(Y^n;\theta),$$
$$\int_{\Theta_0} \pi(\theta) p(Y^n;\theta)\, d\theta \approx C_1 \max_{\theta\in\Theta_1} p(Y^n;\theta).$$

Подставляя эти формулы в байесовский тест приходит к тесту максимального правдоподобия

$$\bar{\phi}_t(Y^n) = \mathbf{1}\Big\{\max_{\theta\in\Theta_1} p(Y^n;\theta) \geq t_\alpha \max_{\theta\in\Theta_0} p(Y^n;\theta)\Big\};$$

здесь параметр t_α подбирается индивидуально в каждом конкретном случае так, чтобы гарантировать заданную вероятность ошибки первого рода

$$\max_{\theta\in\Theta_0} \mathbf{E}_\theta \bar{\phi}_{t_\alpha}(Y^n) \leq \alpha.$$

Несмотря на довольно эвристические аргументы, которые использовались при мотивации теста максимального правдоподобия, этот метод, как правило, дает очень мощный тест. Частные примеры таких тестов будут рассмотрены далее. Оптимальность тестов максимального правдоподобия обычно доказывается с помощью неравенства Фано.

5.3 Неравенство Фано

Оно играет ту же роль в задачах проверки гипотез, что и неравенство Ван Триса в задачах оценивания параметров, ограничивая снизу усредненную вероятность ошибки. Предположим, что наблюдается вектор $Y^n \in \mathbb{R}^n$, плотность распределения которого может быть одной из плотностей $\{p_1(y),\dots,p_M(y)\}$. Задача состоит в том, чтобы по Y^n сказать какова плотность распределения этого вектора. Решением, естественно является некоторая функция $\varphi(Y^n)$, принимающая значения $\{1,\dots,n\}$. Качество этой функции будем измерять средней вероятностью ошибки

$$P_\varphi = 1 - \sum_{k=1}^{M} \pi_k \mathbf{P}_k\big\{\varphi(Y^n) = k\big\};$$

здесь $\pi_k > 0$ — заданные числа, такие что

$$\sum_{k=1}^{M} \pi_k = 1,$$

а

$$\mathbf{P}_k\big\{\varphi(Y^n) = k\big\} = \int_{\mathbb{R}^n} \mathbf{1}\{\varphi(x) = k\} p_k(x)\, dx.$$

Таким образом, мы рассматриваем байесовскую постановку задачи проверки гипотез. При этом считаем, что мы оцениваем случайный параметр θ, принимающий значения $\{1, \ldots, M\}$ с вероятностями $\{\pi_1, \ldots, \pi_M\}$.

Обозначим

$$h(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x},\ x \in [0,1]$$

и определим условную энтропию как

$$H(Y^n) = \sum_{k=1}^{M} \mathbf{P}\{\theta = k | Y^n\} \log \frac{1}{\mathbf{P}\{\theta = k | Y^n\}};$$

здесь

$$\mathbf{P}\{\theta = i | Y^n\} = \pi_i p_i(Y^n) \Big/ \sum_{k=1}^{M} \pi_k p_k(Y^n).$$

Теорема 5.3.1 *Для любого теста $\varphi(\cdot)$ справедливо неравенство*

$$h[P_\varphi] + P_\varphi \log(M-1) \geq \mathbf{E} H(Y^n).$$

Доказательство. Обозначим для краткости

$$q_\varphi(Y^n) = \sum_{i \neq \varphi(Y^n)} \mathbf{P}\{\theta = i | Y^n\} = 1 - \mathbf{P}\{\theta = \varphi(Y^n) | Y^n\}$$

и представим условную энтропию в следующем виде:

$$\begin{aligned} H(Y^n) &= \sum_{i \neq \varphi(Y^n)}^{M} \mathbf{P}\{\theta = i | Y^n\} \log \frac{1}{\mathbf{P}\{\theta = i | Y^n\}} \\ &\quad + P\{\theta = \varphi(Y^n) | Y^n\} \log \frac{1}{\mathbf{P}\{\theta = \varphi(Y^n) | Y^n\}} \\ &= q_\phi(Y^n) \sum_{i \neq \varphi(Y^n)}^{M} \frac{\mathbf{P}\{\theta = i | Y^n\}}{q_\phi(Y^n)} \log \frac{q_\phi(Y^n)}{\mathbf{P}\{\theta = i | Y^n\}} \\ &\quad + \mathbf{P}\{\theta = \varphi(Y^n) | Y^n\} \log \frac{1}{\mathbf{P}\{\theta = \varphi(Y^n) | Y^n\}} + q_\phi(Y^n) \log \frac{1}{q_\phi(Y^n)}. \end{aligned}$$

Из неравенства Йенсена получаем

$$\sum_{i\neq\varphi(Y^n)}^{M}\frac{\mathbf{P}\{\theta=i|Y^n\}}{q_\phi(Y^n)}\log\frac{q_\phi(Y^n)}{\mathbf{P}\{\theta=i|Y^n\}}\le\log(M-1).$$

Заметим далее, что

$$\mathbf{P}\{\theta=\varphi(Y^n)|Y^n\}\log\frac{1}{\mathbf{P}\{\theta=\varphi(Y^n)|Y^n\}}+q_\phi(Y^n)\log\frac{1}{q_\phi(Y^n)}=h[q_\varphi(Y^n)].$$

Следовательно, приходим к следующему неравенству:

$$H(Y^n)\le q_\varphi(Y^n)\log(M-1)+h[q_\varphi(Y^n)].$$

Поскольку $h(x)$ вогнутая функция, и $\mathbf{E}q_\varphi(Y^n)=P_\varphi$, из неравенства Йенсена получаем

$$\begin{aligned}\mathbf{E}H(Y^n)\le&\mathbf{E}q_\varphi(Y^n)\log(M-1)+h[\mathbf{E}q_\varphi(Y^n)]\\=&P_\varphi\log(M-1)+h(P_\varphi).\quad\blacksquare\end{aligned}$$

5.4 Некоторые стандартные тесты

5.4.1 Критерии согласия

Гауссовские распределения

Предположим, что у нас имеется выборка $Y^n=\{Y_1,\dots,Y_n\}$ из гауссовского закона с параметрами μ, σ^2. Задача состоит в том, чтобы проверить простую гипотезу

$$H_0:\ \mu=\mu_0$$

против сложной альтернативы

$$H_0:\ \mu\neq\mu_0,$$

где μ_0 — заданная величина.

Обозначим

$$\begin{aligned}L(\mu,\sigma^2;Y^n)=&\log[p(Y^n,\mu,\sigma^2)]=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2-\frac{n}{2}\log(2\pi\sigma^2)\\=&-\frac{\|Y^n-\mu\|^2}{2\sigma^2}-\frac{n}{2}\log(2\pi\sigma^2)\end{aligned}$$

логарифм правдоподобия для наблюдений Y^n. Тогда тест максимального правдоподобия имеет вид

$$\varphi(Y^n) = \mathbf{1}\Big\{\max_{\sigma^2} L(\mu_0, \sigma^2; Y^n) - \max_{\mu,\sigma^2} L(\mu, \sigma^2; Y^n) \geq t_\alpha\Big\};$$

здесь порог t выбирается так, чтобы гарантировать заданную вероятность ошибки первого рода

$$\mathbf{P}_{\mu_0}\Big\{\max_{\sigma^2} L(\mu_0, \sigma^2; Y^n) - \max_{\mu,\sigma^2} L(\mu, \sigma^2; Y^n) \geq t_\alpha\Big\} = \alpha.$$

Это — безусловно, формальный ответ в рассматриваемой задаче. Наш следующий шаг состоит в том, чтобы его конкретизировать, вычислив соответствующие максимумы. Нетрудно проверить, что

$$\max_{\sigma^2} L(\mu, \sigma^2; Y^n) = -\frac{n}{2} - \frac{n}{2}\log\frac{2\pi\|Y^n - \mu\|^2}{n}$$

и

$$\max_{\mu,\ \sigma^2} L(\mu, \sigma^2; Y^n) = -\frac{n}{2} - \frac{n}{2}\log\frac{2\pi\|Y^n - \bar{Y}^n\|^2}{n}, \tag{5.4.4}$$

где

$$\bar{Y}^n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Таким образом, тест максимального правдоподобия принимает следующий вид:

$$\phi(Y^n) = \mathbf{1}\Big\{\frac{\|Y^n - \mu_0\|^2}{\|Y^n - \bar{Y}^n\|^2} \geq t'_\alpha\Big\}.$$

Чтобы упростить еще этот тест, заметим

$$\|Y^n - \mu_0\|^2 = \|Y^n - \bar{Y}^n + \bar{Y}^n - \mu_0\|^2 = \|Y^n - \bar{Y}^n\|^2 + n|\bar{Y}^n - \mu_0|^2.$$

Следовательно, мы приходим к следующему тесту:

$$\phi(Y^n) = \mathbf{1}\Big\{\frac{|\bar{Y}^n - \mu_0|\sqrt{n}}{\|Y^n - \bar{Y}^n\|/\sqrt{n-1}} \geq t_\alpha\Big\}.$$

Чтобы найти порог t_α нам, естественно, нужно знать распределение статистики

$$T(Y^n) = \frac{|\bar{Y}^n - \mu_0|\sqrt{n}}{\|Y^n - \bar{Y}^n\|/\sqrt{n-1}}$$

при гипотезе, т.е. когда $Y_i = \mu_0 + \sigma\xi_i$, где ξ_i — стандартный белый шум.

Мы уже видели (теорема 3.2.2), что:

- случайные величины $|Y^n - \mu_0|\sqrt{n}$ и $\|Y^n - \bar{X}^n\|/\sqrt{n-1}$ независимы;
- случайная величина $(Y^n - \mu_0)\sqrt{n}$ имеет гауссовское распределение с нулевым средним и дисперсией σ^2;
- случайная величина $\|Y^n - \bar{Y}^n\|^2/\sigma^2$ имеет распределение χ-квадрат с $n-1$ степенью свободы.

Таким образом, статистика $T(Y^n)$ распределена также как $|\eta_{n-1}|$, где

$$\eta_k = \xi_0 \Big/ \left[\frac{1}{k}\sum_{i=1}^{k} \xi_i^2\right]^{1/2};$$

здесь $\xi_i,\ i = 0, \ldots, k$ — независимые, стандартные гауссовские случайные величины.

Распределение случайной величины η_k называется *распределением Стьюдента* с k степенями свободы.

Дискретные распределения

Говорят, что случайная величина Y дискретна, если она принимает только конечное число значений. Сами эти значения роли не играют, поэтому для простоты будем считать, что это целые числа $1, 2 \ldots, d$. Дискретное распределение Y задается числами

$$p(k) = \mathbf{P}\big\{Y = k\big\}, \quad k = 1, \ldots, d.$$

Рассмотрим задачу проверки по выборке $Y^n = \{Y_1, \ldots, Y_n\}$ простой гипотезы

$$H_0 :\ p = p_0$$

против сложной альтернативы

$$H_1 :\ p \neq p_0.$$

Здесь равенство $p = p_0$ означает, что $p(k) = p_0(k)$ для всех $k = 1, \ldots, d$, а $p \neq p_0$ — что существует по крайней мере одно k, для которого $p(k) \neq p_0(k)$. Предполагается также, что величины $p_0(k)$ заданы.

Тест максимального правдоподобия имеет следующий вид

$$\varphi(Y^n) = \mathbf{1}\left\{\max_p \sum_{i=1}^{n} \log[p(Y_i)] - \sum_{i=1}^{n} \log[p_0(Y_i)] \geq t\right\}.$$

Упростим этот тест, обозначив для краткости

$$\bar{p}(k) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i = k\}.$$

Величины $\bar{p}(k)$ являются эмпирическими версиями (оценками) неизвестных $p(k)$.

Тогда очевидно, что

$$\sum_{i=1}^{n} \log[p_0(Y_i)] = n\sum_{k=1}^{d} \bar{p}(k)\log[p_0(k)],$$

а

$$\max_p \sum_{i=1}^{n} \log[p(Y_i)] = n \max_p \sum_{k=1}^{d} \bar{p}(k)\log[p(k)],$$

Вычисления максимума по p является несложной задачей. Нетрудно проверить, что он достигается при $p(k) = \bar{p}(k)$. Действительно, в силу неравенства Йенсена и выпуклости функции $-\log(x)$ имеем для любых $p(k)$

$$\sum_{k=1}^{d} \bar{p}(k)\log[\bar{p}(k)] - \sum_{k=1}^{d} \bar{p}(k)\log[p(k)] = -\sum_{k=1}^{d} \bar{p}(k)\log\frac{p(k)}{\bar{p}(k)} \geq 0.$$

Таким образом, тест максимального правдоподобия имеет следующий вид:

$$\varphi(Y^n) = \mathbf{1}\left\{ n\sum_{k=1}^{d} \bar{p}(k)\log\frac{\bar{p}(k)}{p_0(k)} \geq t_\alpha \right\}.$$

Его вероятностный смысл очень прозрачен: статистика

$$K(Y^n) = n\sum_{k=1}^{d} \bar{p}(k)\log\frac{\bar{p}(k)}{p_0(k)}$$

это эмпирическая версия расстояния Кульбака-Лейблера между p и p_0. Если это расстояние мало то, гипотеза H_0 принимается, а в противном случае — отвергается.

Величину порога t, которая определяется как корень уравнения

$$\mathbf{P}_{H_0}\left\{ n\sum_{k=1}^{d} \bar{p}(k)\log\frac{\bar{p}(k)}{p_0(k)} \geq t_\alpha \right\} = \alpha,$$

можно просто вычислить методом Монте-Карло, поскольку гипотеза является простой.

Другой способ вычисления t_α связан с асимптотической при $n \to \infty$ аппроксимацией статистики $K(Y^n)$. Замечая, что при гипотезе и $n \to \infty$ в силу закона больших чисел $p(k) \to p_0(k)$. Поэтому, пользуясь формулой Тейлора, найдем

$$\log \frac{\bar{p}(k)}{p_0(k)} = -\log\left[1 + \frac{p_0(k)}{\bar{p}(k)} - 1\right] \approx -\frac{p_0(k)}{\bar{p}(k)} + 1 + \frac{1}{2}\left[\frac{p_0(k)}{\bar{p}(k)} - 1\right]^2.$$

Таким образом, получаем следующую аппроксимацию:

$$K(Y^n) \approx \frac{1}{2}\chi^2(Y^n) = \frac{n}{2}\sum_{i=1}^{d} \frac{[\bar{p}(k) - p_0(k)]^2}{p_0(k)}.$$

В традиционной литературе статистика $\chi^2(Y^n)$ называется статистикой χ-квадрат, а тест $\mathbf{1}\{\chi^2(Y^n) > t_\alpha\}$ — тестом χ-квадрат.

Асимптотическое распределение статистики χ-квадрат легко вычисляется.

Теорема 5.4.1 *При $n \to \infty$*

$$\lim_{n\to\infty} \mathbf{P}_{H_0}\{\chi^2(Y^n) \le x\} = \mathbf{P}\left\{\sum_{i=1}^{d-1} \xi_i^2 \le x\right\},$$

где ξ_i — независимые, стандартные гауссовские случайные величины.

Доказательство. Рассмотрим случайные величины

$$\zeta_k^n = \frac{\bar{p}(k) - p_0(k)}{\sqrt{p_0(k)}}\sqrt{n}.$$

Очевидно, что $\mathbf{E}\zeta_k^n = 0$. Кроме того, легко проверить, что

$$\begin{aligned}
\mathbf{E}\zeta_k^n\zeta_j^n = \frac{1}{\sqrt{p_0(k)p_0(j)}}\mathbf{E}\frac{1}{n}\sum_{i,s=1}^{n}\big[\mathbf{1}\{Y_i = j\} - \mathbf{E}\mathbf{1}\{Y_i = j\}\big] \\
\times\big[\mathbf{1}\{Y_s = k\} - \mathbf{E}\mathbf{1}\{Y_s = k\}\big] \\
= \mathbf{E}\big[\mathbf{1}\{Y_1 = j\} - \mathbf{E}\mathbf{1}\{Y_1 = j\}\big] \times \big[\mathbf{1}\{Y_1 = k\} - \mathbf{E}\mathbf{1}\{Y_1 = k\}\big] \\
/\sqrt{p_0(k)p_0(j)} \\
= \delta_{kj} - \sqrt{p_0(k)p_0(j)}.
\end{aligned}$$

Здесь

$$\delta_{kj} = \begin{cases} 1, & j = k, \\ 0, & j \neq k \end{cases}$$

— символ Кронекера.

Поэтому, согласно центральной предельной теореме,

$$(\zeta_1^n, \ldots, \zeta_d^n) \xrightarrow{\mathbf{P}} (\eta_1, \ldots, \eta_d)$$

и

$$\sum_{s=1}^{d} (\zeta_s^n)^2 \xrightarrow{\mathbf{P}} \sum_{s=1}^{d} \eta_s^2,$$

где $\eta_l,\ l = 1, \ldots, d$ — гауссовские случайные величины с корреляционной функцией

$$\mathbf{E}\eta_k\eta_j = \delta_{kj} - \sqrt{p_0(k)p_0(j)}.$$

Обозначим для краткости

$$q_0 = \left(\sqrt{p_0(1)}, \ldots, \sqrt{p_0(1)}\right)^\top$$

и $\varepsilon = \big(\varepsilon_1, \ldots, \varepsilon_d\big)^\top$, где ε_s — независимые, стандартные гауссовские случайные величины.

Рассмотрим гауссовский случайный вектор

$$\xi = \varepsilon - \langle \varepsilon, q_0 \rangle q_0.$$

Легко проверить, что

$$\mathbf{E}\xi_k\xi_j = \mathbf{E}\left[\varepsilon_k - \sqrt{p_0(k)}\sum_{i=1}^{d} \varepsilon_i \sqrt{p_0(i)}\right]\left[\varepsilon_j - \sqrt{p_0(j)}\sum_{i=1}^{d} \varepsilon_i \sqrt{p_0(i)}\right] = \delta_{jk} - \sqrt{p_0(j)p_0(k)}.$$

Таким образом, $\eta_k = \varepsilon_k - \langle \varepsilon, q_0 \rangle q_0$ и, следовательно,

$$\|\eta\|^2 = \min_y \|\varepsilon - yq_0\|^2 = \sum_{i=1}^{d-1} \langle \varepsilon, q_i \rangle^2,$$

где $\big\{q_0, q_1, \ldots, q_{d-1}\big\}$ — ортонормальная система векторов в $\mathbf{R}^d$. Заметив, что $\langle \varepsilon, q_i \rangle$ — независимые, стандартные гауссовские случайные величины, завершаем доказательство теоремы. ■

5.4.2 Критерии сравнения

Гауссовские наблюдения

Предположим, что имеются две гауссовские выборки $Y^n = \{Y_1, \dots, Y_n\}$ и $Z^m = \{Z_1, \dots, Z_m\}$ с параметрами (μ_Y, σ_Y^2) и (μ_Z, σ_Z^2) соответственно. Задача состоит в том, чтобы по этим наблюдениям проверить гипотезу

$$H_0 : \ \mu_Y = \mu_Y, \ \sigma_Y^2 = \sigma_Y^2$$

против альтернативы

$$H_1 : \ (\mu_Y, \sigma_Y^2) \neq (\mu_Z, \sigma_Z^2).$$

Для построения теста максимального правдоподобия воспользуемся формулой (5.4.4). Из нее легко видеть, что этот тест имеет следующий вид:

$$\varphi(Y^n, Z^m) = \mathbf{1}\big\{T(Y^n, Z^m) > t\big\},$$

где

$$T(Y^n, Z^m) = -\frac{m+n}{2} - \frac{m+n}{2} \log\left[\min_{\mu} \frac{\|Y^n - \mu\|^2 + \|Z^m - \mu\|^2}{n+m}\right] + \frac{n}{2} + \frac{n}{2} \log \frac{\|Y^n - \bar{Y}^n\|^2}{n} + \frac{m}{2} + \frac{m}{2} \log \frac{\|Z^m - \bar{Z}^m\|^2}{m}.$$

Заметим, значение μ^*, на котором достигается

$$\min_{\mu} \frac{\|Y^n - \mu\|^2 + \|Z^m - \mu\|^2}{n+m},$$

имеет вид

$$\mu^* = \frac{n\bar{Y}^n + m\bar{Z}^m}{n+m}.$$

Поэтому

$$\min_{\mu} \frac{\|Y^n - \mu\|^2 + \|Z^m - \mu\|^2}{n+m} = \frac{\|Y^n - \bar{Y}^n\|^2 + \|Z^m - \bar{Z}^m\|^2}{m+n} + \frac{mn|\bar{Y}^n - \bar{Z}^m|^2}{(m+n)^2}$$

и, следовательно, статистика теста вычисляется следующим образом:

$$T(Y^n, Z^m) = \frac{n}{2} \log \frac{\|Y^n - \bar{Y}^n\|^2}{n} + \frac{m}{2} \log \frac{\|Z^m - \bar{Z}^m\|^2}{m} - \frac{m+n}{2} \log\left[\frac{\|Y^n - \bar{Y}^n\|^2 + \|Y^m - \bar{Y}^m\|^2}{m+n} + \frac{mn|\bar{Y}^n - \bar{Z}^m|^2}{(m+n)^2}\right]$$

Нетрудно видеть, что при гипотезе распределение этой случайной величины не зависит ни от $\mu_Y = \mu_Z$, ни от $\sigma_Y^2 = \sigma_Z^2$. Поэтому критическое значение t_α для этой статистики, которое является корнем уравнения

$$\mathbf{P}_{H_0}\{T(Y^n, Z^m) > t_\alpha\} = \alpha \tag{5.4.5}$$

можно легко вычислить методом Монте-Карло, взяв $\mu_X = \mu_Y = 0$ и $\sigma_X^2 = \sigma_Y^2 = 1$. Точнее, пусть $\{\xi_0, \xi_1, \ldots, \xi_{n-1}, \xi_n, \ldots, \xi_{n+m-2}\}$ — независимые, стандартные гауссовские случайные величины. Тогда, как мы уже видели, при гипотезе справедливо соотношение

$$T(Y^n, Y^m) \overset{\mathbf{P}_{H_0}}{=} \frac{n}{2}\log\left[\frac{1}{n}\sum_{i=1}^{n-1}\xi_i^2\right] + \frac{m}{2}\log\left[\frac{1}{m}\sum_{i=n}^{n+m-2}\xi_i^2\right] - \frac{m+n}{2}\log\left[\frac{1}{m+n}\sum_{i=0}^{n+m-1}\xi_i^2\right].$$

Поэтому, для того чтобы найти корень уравнения (5.4.5), достаточно вычислить методом Монте-Карло функцию распределения правой части в этом тождестве.

Тестирование аспирина

В этом разделе, чтобы проиллюстрировать как используются тесты на практике, рассмотрим с точке зрения теории статистических тестов широко распространенную задачу тестирования медикаментов. В частности речь пойдет о тестировании аспирина, которое было описано в статье *Heart attack risk found to be cut by taking aspirin*, появившейся в New York Times 27 января 1997.

Задача состоит с в том, чтобы выяснить какое влияние оказывает прием аспирина на вероятность сердечного приступа. Для этого были сформированы две группы пациентов. Пациенты первой группы принимали аспирин (11037 пациентов), а пациенты второй группы (11034 пациентов) — плацебо. Врачи, наблюдавшие пациентов, не знали что принимают их пациенты: аспирин или плацебо. Само собой разумеется, что и пациенты этого не знали. В каждой группе регистрировалось число сердечных приступов, данные о которых сведены в следующую таблицу:

	число сердечных приступов	число пациентов
аспирин	104	11037
плацебо	189	11034

Вероятностная модель для интерпретации этих данных очень проста. Каждому пациенту припишем двоичное число 0 или 1

- 1 говорит о том, что у пациента был сердечный приступ,
- 0 — что нет.

Таким образом, в качестве наблюдений у нас имеются два бинарных вектора

$$A^m = \{A_1, \ldots, A_m\},\ m = 11037$$
$$P^n = \{P_1, \ldots, P_n\},\ n = 11034.$$

Вектор A^n описывает группу пациентов, принимающих аспирин, а P^m — плацебо. Довольно естественно предполагать, что сами векторы и их компоненты — независимые, одинаково распределенные бернулиевские случайные величины:

$$\mathbf{P}\{A_i = 1\} = p_a, \quad \mathbf{P}\{A_i = 0\} = 1 - p_a,$$
$$\mathbf{P}\{P_i = 1\} = p_p, \quad \mathbf{P}\{P_i = 0\} = 1 - p_p.$$

Таким образом, мы приходим к задаче проверки по наблюдениям $\{A^m, P^n\}$ гипотезы о том, что аспирин действует также как плацебо

$$H_0 : \quad p_a = p_p$$

против сложной альтернативы, что аспирин уменьшает вероятность сердечного приступа

$$H_1 : \quad p_a \neq p_p.$$

Как и ранее для решения этой задачи будем использовать тест максимального правдоподобия. Обозначим для краткости

$$\tilde{A} = \sum_{i=1}^{m} A_i, \quad \tilde{P} = \sum_{i=1}^{n} P_i$$

Эти величины обозначают числа сердечный приступов в группах пациентов, принимающих аспирин и плацебо. Заметим, что вероятность того, что в бинарном векторе $B^k = (B_1, \ldots, B_k)$ с независимыми компонентами содержится j единиц вычисляется как

$$\mathbf{P}\left\{\sum_{s=1}^{k} B_s = j\right\} = C_k^j p^j (1-p)^{k-j}$$

где

$$C_k^j = \frac{k!}{(k-j)!j!}, \quad p = \mathbf{P}\{B_s = 1\}.$$

Это распределение называется *распределением Бернулли*.

Поэтому тест максимального правдоподобия имеет следующий вид:

$$\begin{aligned}\varphi(A^m, P^n) =& \mathbf{1}\Big\{\max_p\Big[(\tilde{A}+\tilde{P})\log(p) + (m+n-\tilde{A}-\tilde{P})\log(1-p)\Big] \\ &- \max_{p,q}\Big[\tilde{A}\log(p) + (m-\tilde{A})\log(1-p) \\ &+ \tilde{P}\log(q) + (n-\tilde{P})\log(1-q)\Big] > t_\alpha\Big\}.\end{aligned}$$

Чтобы упростить этот тест заметим, что

$$\max_p\Big[x\log(p) + y\log(1-p)\Big] = -(x+y)h\left(\frac{x}{x+y}\right),$$

где энтропия $h(x),\ x \in [0,1]$ определяется стандартным образом

$$h(x) = -x\log(x) - (1-x)\log(1-x).$$

При этом максимум по p достигается при $p = x/(x+y)$.

Поэтому тест максимального принимает следующий вид

$$\varphi(A^m, P^n) = \mathbf{1}\left\{mh\left(\frac{\tilde{A}}{m}\right) + nh\left(\frac{\tilde{P}}{n}\right) - (m+n)h\left(\frac{\tilde{A}+\tilde{P}}{m+n}\right) > t_\alpha\right\}.$$

Порог t_α можно вычислить методом Монте-Карло. Для этого поступим следующим образом:

- оцениваем по имеющимся данным величину p_p

$$\bar{p}_p = \frac{189}{11034} = 0,0171;$$

- генерируем достаточно большое число N (порядка 10000) независимых бинарных векторов $\{A_i^m, P_i^n,\ i = 1, \ldots, N\}$ с независимыми компонентами и вероятностью появления 1 равной $\bar{p}_p$;

- для каждой пары $\{A_i^m, P_i^n\}$ вычисляем тестовую статистику

$$T_i = mh\left(\frac{\tilde{A}_i}{m}\right) + nh\left(\frac{\tilde{P}_i}{n}\right) - (m+n)h\left(\frac{\tilde{A}_i+\tilde{P}_i}{m+n}\right),$$

где

$$\tilde{A}_i = \sum_{k=1}^{m} A_{ik}^m, \quad \tilde{P}_i = \sum_{k=1}^{n} P_{ik}^n;$$

- упорядочиваем величины $T_i, i = 1, \ldots, N$ и в качестве порога выбираем

$$t_\alpha = T_{(\alpha n)}$$

Глава 6

Линейные модели

6.1 Метод наименьших квадратов

Ранее достаточно подробно была изучена статистическая модель оценивания параметра сдвига $\mu \in \mathbb{R}$ по наблюдениям

$$Y_i = \mu + \epsilon_i, \quad i = 1, \dots, n.$$

В этой модели все Y_i являются одинаково распределенными случайными величинами. Одинаковая распределенность наблюдений скорее является скорее исключением, чем правилом. Во многих случаях просто невозможно обойтись одинаково распределенными данными, чтобы смоделировать ту или иную реальную ситуацию. Как правило, мы наблюдаем пары случайных величин (X_i, Y_i), $i = 1, \dots, n$ и интересует как величина Y_i зависит от X_i. Например, если X_i является площадью квартиры, а Y_i ее ценой, то было бы интересно знать какова зависимость между этими величинами. Чтобы решить эту задачу мы могли бы предположить, что зависимость между площадью и ценой линейна, т.е.

$$Y_i = \mu_1 + \mu_2 X_i + \epsilon_i, \quad i = 1, \dots, n. \tag{6.1.1}$$

В рамках этой модели, установить зависимость между Y_i и X_i означает оценить по наблюдениям $Y = (Y_1, \dots, Y_n)^\top$ векторный параметр $\mu = (\mu_1, \mu_2)^\top$.

Заметим, что наблюдения удобно представить в следующей матричной форме. Определим матрицу

$$X = \begin{pmatrix} 1, & 1, \dots & ,1 \\ X_1, & X_2 \dots, & X_n \end{pmatrix}^\top.$$

Тогда мы приходим к самой широко распространенной в статистике линейной модели

$$Y = X\mu + \epsilon; \tag{6.1.2}$$

здесь

- $\mu \in \mathbb{R}^p$ — неизвестный параметр,
- $Y \in \mathbb{R}^n$ — наблюдения,
- X — известная $n \times p$-матрица,
- $\epsilon \in \mathbb{R}^n$ – шум с известной или неизвестной интенсивностью σ (ϵ_i — независимые случайные величины с нулевым средним $\mathbf{E}\epsilon_i = 0$ и конечной дисперсией $\mathbf{E}\epsilon_i^2 = \sigma^2$).

В рассматриваемом случае линейной зависимости размерность оцениваемого параметра $p = 2$. Конечно, простая гипотеза о линейной зависимости между площадью и ценой является достаточно наивной. Кроме площади есть еще несколько существенных параметров, которые влияют на стоимость квартиры, например, такие как близость к метро, район, этаж. Все их легко добавить в линейную модель, расширив размерность матрицы X и соответственно вектора μ.

Оценка наименьших квадратов параметра μ в модели (6.1.2) определяется как

$$\hat{\mu}_\circ(Y) = \arg\min_\mu \|Y - X\mu\|^2$$

или как корень уравнения

$$X^\top X\hat{\mu}_\circ = X^\top Y.$$

Отсюда легко видеть, что если матрица $X^\top X$ невырождена, то

$$\hat{\mu}_\circ(Y) = (X^\top X)^{-1}X^\top Y = \mu + (X^\top X)^{-1}X^\top\epsilon. \tag{6.1.3}$$

Для статистического анализа этой оценки удобно использовать метод главных компонент. Обозначим e_k и λ_k собственные векторы и собственные числа матрицы $X^\top X$:

$$X^\top Xe_k = \lambda_k e_k, \quad k = 1, \ldots, p.$$

(Будем считать для определенности, что $\lambda_1 \geq \ldots \geq \lambda_p > 0$.)

Тогда легко видеть, что векторы

$$e_k^* = \frac{Xe_k}{\sqrt{\lambda_k}}$$

ортонормальны

$$\langle e_k^*, e_j^* \rangle = \frac{\langle Xe_k, Xe_j \rangle}{\sqrt{\lambda_k \lambda_j}} = \frac{\langle X^\top X e_k, e_j \rangle}{\sqrt{\lambda_k \lambda_j}} = \frac{\lambda_k \langle e_k, e_j \rangle}{\sqrt{\lambda_k \lambda_j}}.$$

Рассмотрим два линейных преобразования наблюдений Y

$$\begin{aligned} \tilde{Y}_k^* &= \langle Y, e_k^* \rangle = \langle X\mu, e_k^* \rangle + \sigma \xi_k', \\ \tilde{Y}_k &= \frac{\langle X^\top Y, e_k \rangle}{\lambda_k} = \langle \mu, e_k \rangle + \frac{\sigma}{\sqrt{\lambda_k}} \xi_k'; \end{aligned} \tag{6.1.4}$$

здесь ξ_k' – гауссовские независимые с.в. $\mathcal{N}(0,1)$.

Поэтому, используя для оценивания $X\mu$ вектор $\tilde{Y}^*$, а для оценивания μ — вектор $\tilde{Y}$, сразу же получаем следующую теорему.

Теорема 6.1.1 *Если $X^\top X > 0$, то*

$$\mathbf{E}\|\hat{\mu}_\circ - \mu\|^2 = \sigma^2 \sum_{k=1}^{p} \frac{1}{\lambda_k}, \quad \mathbf{E}\|X\hat{\mu}_\circ - X\mu\|^2 = \sigma^2 p.$$

Следующая теорема резюмирует статистические свойства метода максимального правдоподобия в случае гауссовских ошибок ϵ в линейной модели.

Теорема 6.1.2 *Если $X^\top X > 0$, то*

1. *случайный вектор $(\hat{\mu}_\circ - \mu)/\sigma$ является гауссовским с нулевым средним и матрицей ковариации $B = (X^\top X)^{-1}$;*
2. *случайная величина $\|X\hat{\mu}_\circ - X\mu\|^2/\sigma^2$ имеет стандартное χ - квадрат распределение с p степенями свободы;*
3. *случайная величина $\|X\hat{\mu}_\circ - Y\|^2/\sigma^2$ имеет стандартное χ - квадрат распределение с $n-p$ степенями свободы.*

Доказательство. Первое утверждение теоремы вытекает практически непосредственно из (6.1.3) если заметить, что

$$\begin{aligned} \mathbf{E}(\hat{\mu}_\circ - \mu)(\hat{\mu}_\circ - \mu)^\top &= \sigma^2 \mathbf{E}(X^\top X)^{-1} X^\top \epsilon \big[(X^\top X)^{-1} X^\top \epsilon\big]^\top \\ &= \sigma^2 (X^\top X)^{-1} X^\top \mathbf{E}\epsilon\epsilon^\top X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}. \end{aligned}$$

Утверждения 2 и 3 вытекают из геометрического смысла величины

$$\|X\hat{\mu}_\circ - Y\|^2 = \min_\mu \|X\mu - Y\|^2,$$

которая представляет собой квадрат проекции вектора Y на подпространство ортогональное подпространству натянутому на векторы, образованные из столбцов матрицы X. ■

Из теоремы 6.1.1 вытекает, в частности, что несмещенная оценка дисперсии шума σ^2 имеет вид

$$\bar{\sigma}^2(Y) = \frac{\|X\hat{\mu}_\circ - Y\|^2}{n-p}.$$

Числом обусловленности матрицы X называется корень из отношение максимального собственного числа $X^\top X$ к минимальному

$$\mathrm{cond}(X) \stackrel{\mathrm{def}}{=} \sqrt{\frac{\lambda_1}{\lambda_p}}.$$

Если число обусловленности матрицы X велико или размерность оцениваемого параметра p велика, то чтобы уменьшить риск необходимо использовать априорную информацию об оцениваемом векторе μ. В этом случае говорят о регуляризации метода наименьших квадратов, которую мы рассмотрим в следующем разделе.

6.2 Простейшие методы регуляризации

Регуляризация Тихонова (ridge regression)

Предположим, что μ – гауссовский случайный вектор с независимыми $\mathcal{N}(0, 1/\alpha)$ компонентами. В этом случае байессовская оценка имеет вид

$$\hat{\mu}_\alpha = \arg\min_\mu \left\{ \frac{\|Y - X\mu\|^2}{2\sigma^2} + \frac{\alpha}{2}\|\mu\|^2 \right\}.$$

Решение этой оптимизационной задачи легко находится

$$\hat{\mu}_\alpha = [X^\top X + \sigma^2 SI]^{-1} X^\top Y = H_\alpha[X^\top X]\hat{\mu}_\circ,$$

где

$$H_\alpha[X^\top X] = [X^\top X + \alpha\sigma^2 I]^{-1} X^\top X,$$

а I — единичная матрица. При практической реализации этого метода никакие обратные матрицы, естественно, не вычисляются. Оценка $\hat{\mu}_\alpha$ находится как решение системы линейных уравнений

$$[X^\top X + \sigma^2 SI]\hat{\mu}_\alpha = X^\top Y.$$

Заметим, что для матрицы регуляризации в методе Тихонова справедливо спектральное представление

$$H_\alpha[X^\top X] = \sum_{k=1}^{p} H_\alpha(\lambda_k) e_k e_k^\top,$$

где

$$H_\alpha(\lambda) = \frac{\lambda}{\lambda + \alpha\sigma^2}.$$

Метод Ландвебера

Этот метод основан на простой идее: решить рекуррентным методом уравнение

$$X^\top X\mu = X^\top Y.$$

Поскольку

$$X^\top Y = \big[X^\top X - \gamma I\big]\mu + \gamma\mu$$

для всех $\alpha > 0$, то имеем

$$\mu = \big[I - \gamma^{-1}X^\top X\big]\mu + \gamma^{-1}X^\top Y.$$

Поэтому мы можем вычислять корень как

$$\hat\mu^{(k)} = \big[I - \gamma^{-1}X^\top X\big]\hat\mu^{(k-1)} + \gamma^{-1}X^\top Y, \quad \hat\mu^{(0)} = 0. \tag{6.2.5}$$

Таким образом, мы можем оценивать μ без применения SVD и без решения системы линейных уравнений.

Нетрудно проверить, что этот метод сходится при $\gamma > \lambda_1$ и что его регуляризационная матрица имеет следующий вид:

$$H^{(k)}(X^\top X) = I - \big(I - \gamma^{-1}X^\top X\big)^{k+1}. \tag{6.2.6}$$

Действительно, подставив в уравнение (6.2.5)

$$\mu^{(k)} = H^{(k)}(X^\top X)(X^\top X)^{-1}X^\top Y,$$

получим рекуррентное уравнение

$$H^{(k)}(X^\top X) = \big[I - \gamma^{-1}X^\top X\big]H^{(k-1)}(X^\top X) - \gamma^{-1}(X^\top X),$$

решением которого и является (6.2.6).

К сожалению, метод Ландвебера может сходиться очень медленно. Это происходит тогда, когда число обусловленности матрицы X велико. Нетрудно видеть, число итераций метода может быть порядка

$$k \gtrsim \operatorname{cond}^2(X^\top X) = \frac{\lambda(1)}{\lambda(n)}.$$

Действительно, k должно быть таким, чтобы $H^{(k)}(\lambda_n) \approx 1$. Это условие эквивалентно тому, что

$$1 \gg \left(1 - \frac{\lambda_1}{\gamma} \times \frac{\lambda_n}{\lambda_1}\right)^{k+1} \approx \exp\left[-(k+1)\frac{\lambda_1}{\gamma} \times \frac{\lambda_n}{\lambda_1}\right].$$

Spectral cut-off

Для этого метода

$$H_\alpha(\lambda) = \mathbf{1}\{\lambda \geq \alpha\}.$$

Это самый затратный в вычислительном отношении метод регуляризации поскольку он требует вычисления вычисления собственных векторов и собственных чисел матрицы $X^\top X$.

6.2.1 Риск спектральной регуляризации

Методы регуляризации Тихонова, Ландвебера и spectral cut-off являются частными случаем широкого класса спектральных методов регуляризации, которые имеют следующий вид:

$$\hat{\mu}_\alpha = H_\alpha[X^\top X]\hat{\mu}_\circ(Y), \text{ где } H_\alpha[X^\top X] = \sum_{k=1}^{p} H_\alpha(\lambda_k) e_k e_k^\top.$$

Функция $H_\alpha(\lambda)$, как правило, принимает значения из $[0,1]$ и такова, что

$$\lim_{\lambda \to 0} H_\alpha(\lambda) = 0, \quad \lim_{\alpha \to 0} H_\alpha(\lambda) = 1.$$

Риск спектральной регуляризации описывает следующая теорема.

Теорема 6.2.1 *Справедливы соотношения*

$$\mathbf{E}\|\hat{\mu}_\alpha - \mu\|^2 = \sum_{k=1}^{p}[1 - H_\alpha(\lambda_k)]^2\langle\mu, e_k\rangle^2 + \sigma^2 \sum_{k=1}^{p} \frac{H_\alpha^2(\lambda_k)}{\lambda_k},$$

$$\mathbf{E}\|X\hat{\mu}_\alpha - X\mu\|^2 = \sum_{k=1}^{p}\lambda_k[1 - H_\alpha(\lambda_k)]^2\langle\mu, e_k\rangle^2 + \sigma^2 \sum_{k=1}^{p} H_\alpha^2(\lambda_k).$$

Доказательство. Оно вытекает непосредственно из спектрального представления наблюдений (6.1.4). ■

До сих пор мы ничего не говорили про выбор параметра регуляризации α. Естественно выбирать этот параметр так, чтобы минимизировать

$$\text{либо} \quad \mathbf{E}\|\hat{\mu}_\alpha - \mu\|^2, \quad \text{либо} \quad \mathbf{E}\|X\hat{\mu}_\alpha - X\mu\|^2.$$

Заметим, что если для выбора параметра α пытаться использовать теорему (6.2.1), то тогда нужно знать величины $\langle \mu, e_k\rangle^2,\ k = 1, \dots, p$ и поэтому данный подход неприемлем с практической точки зрения.

Поэтому очень часто используется следующий выбор:

$$\bar{\alpha} = \arg\min_\alpha \left\{ \|Y - X\hat{\mu}_\alpha\|^2 + 2\sigma^2 \sum_{i=1}^{p} H_\alpha(\lambda_k) \right\}.$$

Этот подход называется методом минимизации несмещенной оценки риска. Изучение его свойств в принципе не очень сложно, но связано с принципиально другими методами статистического анализа и выходит за рамки этого краткого курса.

6.3 Доверительные множества

Оценивание неизвестного параметра $\theta \in \mathbb{R}^p$ по наблюдениям $Y^n \in \mathbb{R}^n$ с помощью доверительных множеств состоит в поиске таких подмножеств $\Theta^\alpha(Y^n) \subset \mathbb{R}^p$ таких, что неравенство

$$\int_{\mathbb{R}^n} \mathbf{1}\big\{\theta \in \Theta^\alpha(y)\big\} p(y;\theta)\, dy \geq 1 - \alpha \tag{6.3.7}$$

выполняется при всех θ. Здесь

- $p(y;\theta)$ — плотность распределения наблюдений Y^n;
- α — уровень доверия, как правило $\alpha = 0,05$.

Иными словами доверительное множество, это множество, построенное на основе наблюдений и такое, что неизвестный параметр принадлежит ему с вероятностью не меньшей, чем $1 - \alpha$.

Тривиальным доверительным множеством является очевидно все пространство. Поэтому мы, конечно, хотим найти среди всех доверительных множеств, удовлетворяющих (6.3.7), множество имеющее минимальную

меру Лебега. К сожалению, точное решение этой задачи в общем случае очень сложно.

Существенно упростить эту задачу можно если перейти к байесовской постановке. Иными словами, вместо (6.3.7) можно определить доверительное множество с помощью неравенства

$$\int_{\mathbb{R}^p} \pi(\theta) \int_{\mathbb{R}^n} \mathbf{1}\big\{\theta \in \Theta^\alpha(y)\big\} p(y;\theta)\, dy\, d\theta \geq 1-\alpha, \tag{6.3.8}$$

где $\pi(\cdot)$ — апостериорная плотность распределения параметра θ. В этом случае наилучшее доверительное множество в принципе легко найти

$$\Theta^\alpha(Y^n) = \left\{\theta : \pi(\theta) p(Y^n;\theta) \Big/ \int_{\mathbb{R}^1} \pi(t) p(Y^n;t)\, dt \geq 1-\alpha\right\}. \tag{6.3.9}$$

Поскольку часто достаточно трудно выбрать априорную плотность $\pi(\cdot)$, на практике, как правило, используют следующую аппроксимацию доверительного множества из (6.3.9):

$$\Theta^\alpha(Y^n) = \left\{\theta : p(Y^n;\theta) \Big/ \max_{t\in\mathbb{R}^p} p(Y^n;t) \geq 1-\alpha\right\}. \tag{6.3.10}$$

Заметим, что для того, чтобы пользоваться этой формулой необходимо точно знать плотность распределения наблюдений $p(\cdot;\theta)$. Это довольно сильное требование. Обычно точно плотность бывает исключительно редко известна. Более реалистичное предположение состоит в том, что она известна с точностью до некоторого неизвестного параметра $\nu \in \mathbb{R}^q$. Это означает что существуют некоторое ν, такое что плотность распределения имеет вид

$$p(y;\theta,\nu),$$

где $\theta \in \mathbb{R}^p$ — неизвестный параметр, который нас интересует, а $\nu \in \mathcal{N} \subset \mathbb{R}^q$ — также неизвестный параметр, но значение которого для нас не важно. Такие параметры в статистике называются мешающими.

При мешающим параметре доверительное множество имеет следующий вид:

$$\Theta^\alpha(Y^n) = \left\{\theta : \max_{\nu\in\mathcal{N}} p(Y^n;\theta,\nu) \Big/ \max_{\nu\in\mathcal{N}, t\in\mathbb{R}^p} p(Y^n;t,\nu) \geq 1-\alpha\right\}. \tag{6.3.11}$$

Применим теперь этот метод построения доверительных множеств для линейной модели при гауссовских ошибках. В этом случае правдоподобие имеет вид

$$p(Y^n;\mu,\sigma^2) = \exp\left\{-\frac{\|Y^n - X\mu\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)\right\}.$$

Роль мешающего параметра играет дисперсия шума σ^2, а роль основного параметра — вектор μ. Максимизируя правдоподобие по σ^2, найдем

$$\begin{aligned}\max_{\sigma^2} p(Y^n;\mu,\sigma^2) &= \exp\left\{-\frac{n}{2}\log\left[2\pi\mathrm{e}\frac{\|Y^n - X\mu\|^2}{n}\right]\right\}\\ &= \left[2\pi\mathrm{e}\frac{\|Y^n - X\mu\|^2}{n}\right]^{-n/2}.\end{aligned}$$

Поэтому доверительное множество определяется следующим образом:

$$\Theta^\alpha(Y^n) = \left\{\mu : \frac{\|Y^n - X\mu\|^2}{\min_{\tilde{\mu}}\|Y^n - X\tilde{\mu}\|^2} \le (1-\alpha)^{-2/n}\right\}.$$

Заметим, что из определения оценки $\hat{\mu}^\circ(Y^n)$ вытекает следующее соотношение

$$\begin{aligned}\|Y^n - X\mu\|^2 &= \|Y^n - X\hat{\mu}^\circ + X\hat{\mu}^\circ - X\mu\|^2\\ &= \|Y^n - X\hat{\mu}^\circ\|^2 + \|X\hat{\mu}^\circ - X\mu\|^2.\end{aligned}$$

Следовательно,

$$\Theta^\alpha(Y^n) = \left\{\mu : \frac{\|X\hat{\mu}^\circ - X\mu\|^2}{\|Y^n - X\hat{\mu}^\circ\|^2} \le (1-\alpha)^{-2/n} - 1\right\}.$$

Иными словами, это доверительное множество представляет собой эллипсоид с центром точке $\hat{\mu}^\circ(Y^n)$ и главными осями, направленными вдоль собственных векторов матрицы $X^\top X$. При этом величина k-ой оси пропорциональна $1/\sqrt{\lambda_k}$.

Глава 7

Задачи

---

1. Пусть $U_i,\ i = 1, \ldots, n$ — независимые, равномерно распределенные на отрезке $[0, 1]$ случайные величины, а $U_{(i)}$ их порядковые статистики. Найти предельное распределение случайной величины

$$\Delta = \min_{i=1,\ldots,n-1} [U_{(i+1)} - U_{(i)}].$$

---

2. Пусть

$$p_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}, \quad p_2(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\},$$

где $x \in \mathbb{R}^1$. Вычислить расстояние Кульбака-Лейблера между плотностями p_1 и p_2.

---

3. Пусть на отрезке [0,1] наблюдается случайный процесс

$$Y(t) = \theta S(t) + \xi(t),$$

где θ — неизвестный случайный параметр, принимающий значения $\{-1, +1\}$ с равными вероятностями, $S(t)$ — известная функция из $L_2[0, 1]$, а $\xi(t)$ — стандартный белый гауссовский шум с непрерывным временем. По наблюдениям $Y(t),\ t \in [0, 1]$ проверяется гипотеза $H_0 :\ \theta = -1$ против альтернативы $H_1 :\ \theta = 1$. Найти байесовский тест в этой задаче и вычислить его вероятность ошибки.

---

4. Пусть имеются наблюдения

$$Y_k = \beta_1 + X_k\beta_2 + \sigma\xi_k,\ k = 1, \ldots, n,$$

где $\{\beta_1, \beta_2\}$ — неизвестные параметры, X_k — известные числа, ξ_k — стандартный белый гауссовский шум с дискретным временем. Найти оценку максимального правдоподобия параметров $\{\beta_1, \beta_2\}$.

---

5. Стационарный случайный процесс с дискретным временем Y_k определяется рекуррентным соотношением

$$Y_k = \alpha Y_{k-1} + \xi_k,$$

где $|\alpha| < 1$, а ξ_k — стандартный белый гауссовский шум с дискретным временем. Найти спектральную плотность процесса Y_k.

---

6. Пусть наблюдения $Y_k,\ k = 1, \ldots, n$ имеют следующий вид:

$$Y_k = \theta + \sigma\xi_k,$$

где θ — неизвестный случайный параметр, имеющий плотность распределения

$$p(x) = \frac{1}{2\lambda}\exp\left[-\frac{|x|}{\lambda}\right],$$

а ξ_k — стандартный белый гауссовский шум с дискретным временем. Найти оценку, максимизирующую апостериорную плотность распределения параметра θ.

---

7. Пусть $Y_k,\ k = 1, \ldots, n$ — независимые случайные величины, имеющие стандартное экспоненциальное распределение. Воспользовавшись неравенством Чернова, оценить сверху

$$\mathbf{P}\left\{\sum_{k=1}^{n} Y_k \geq x\right\}.$$

---

8. Пусть наблюдается случайный процесс с непрерывным временем

$$Y(t) = \theta \cos(2\pi t + \phi) + \xi(t),\ t \in [0, 1],$$

где $\theta > 0$ — неизвестный параметр, который нас интересует, $\phi \in [0, 2\pi]$ — неизвестный мешающий параметр, $\xi(t)$ — стандартный белый гауссовский шум с непрерывным временем. Найти оценку максимального правдоподобия параметра θ.

Литература

[1] *Боровков А.А.* Математическая статистика. М.:ФИЗМАТЛИТ, 2007.

[2] *Ивченко Г.И., Медведев Ю.И.* Введение в математическую статистику. М.: ЛКИ, 2010.

[3] *Ширяев А.Н.* Вероятность. М.: МЦМО, 2007.

[4] *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning. Springer-Verlag N.Y., 2009.

[5] *Wiener N.* Extrapolation, Interpolation, and Smoothing of Stationary Time Series. New York: Wiley, 1949.

[6] *Колмогоров А. Н.* Стационарные последовательности в гильбертовом пространстве // Бюлл. МГУ. 1941. Т. 2. № 6. С. 3–40.

[7] *Колмогоров А.Н.* Интерполирование и экстраполирование стационарных случайных последовательностей // Изв. АН СССР. 1941. Т. 5. № 1. С. 3–14.

[8] *Ван Трис Г.* Теория оценок, обнаружения и модуляции. Том 1. Москва. Советское радио 1972.

[9] *Landweber, L.* (1951). An iteration formula for Fredholm integral equations of the first kind. // Amer. J. Math. **73** 615–624.

[10] *Engl, H.W., Hanke, M., and Neubauer, A.* (1996). Regularization of Inverse Problems. Mathematics and its Applications, 375. Kluwer Academic Publishers Group. Dordrecht.

Алфавитный указатель