УДК 528.1

А. С. ЯРМОЛЕНКО

ОПРЕДЕЛЕНИЕ ФУНКЦИИ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ОШИБОК ИЗМЕРЕНИЙ ДЛЯ УСТОЙЧИВОГО ОЦЕНИВАНИЯ

Вопросы совершенствования моделей ошибок измерений и их применения в уравнительных вычислениях, отбраковки грубых измерений и построения стохастических моделей требуют учитывать всю доступную информацию об измерениях [5].

Наиболее распространенным в теории математической обработки геодезических измерений является метод наименьших квадратов, который обосновывается предположениями теоремы Гаусса—Маркова. Однако отклонение от условий применимости метода наименьших квадратов приводит к более общим методам получения оценок неизвестных. Эти отклонения вызываются также и неоднородностью выборок, что отмечено в [7].

С этой целью рекомендуется [4] исследовать общий закон распределения ошибок и, исходя из него, разработать метод обработки измерений. В [6] предлагается использовать распределения с более утолщенными «хвостами», учитывающие грубые ошибки с тем, чтобы в последующем их исключить. Алгоритмы, основанные на таком предположении, будут обладать устойчивостью по отношению к грубым ошибкам.

Наличием в выборках грубых ошибок нарушается предположение теоремы Гаусса—Маркова о существовании и определенности дисперсий измерений. Это нарушение аналогично отсутствию информации о точности измерений. В связи с этим при выборе функций плотности распределения ошибок, руководствуясь информацией о точности измерений, целесообразно подобрать такую функцию плотности, при которой достигается максимальная дисперсия в оценке неизвестных параметров и которая тем самым предохраняет от возможно большего риска при определении количества информации Фишера. Идея такого, минимаксного по сути, подхода принадлежит [10]. Однако в этой работе не рассматривается применение ее для конкретных условий, не приводятся соответствующие выводы и результаты. Именно практическому осуществлению выдвинутой идеи и посвящаются настоящие исследования, на основе которых могут разрабатываться методы устойчивого оценивания.

В основу исследований положим известное выражение информации Фишера

$$I=\int((\ln y)')^2 y\,dx, \tag{1}$$

или

$$I=\int\frac{y'^2}{y}\,dx, \tag{2}$$

где y — функция плотности распределения.

Для определения экстремального значения функционала (2) при определенных условиях необходимо определить его дифференциал в смысле Фреше [3]

$$\delta I=\int\frac{2y'\,h'\cdot y-y'^2\cdot h}{y^2}\,dx, \tag{3}$$

где h — приращение функции по Фреше.

Производную в (3) будем рассматривать по определяемому параметру. Выражение (3) можно переписать в виде

$$\delta I = \int \left[\frac{2y'}{y} h' - \left(\frac{y'}{y}\right)^2 h \right] dx. \qquad (4)$$

Для приведения выражения (4) к единому приращению рассмотрим следующую производную по параметру t:

$$\left(\frac{2y'}{y} h\right)' = h \frac{d}{dt}\left(\frac{2y'}{y}\right) + \frac{2y'}{y} h'. \qquad (5)$$

Решая (5) относительно $\frac{2y'}{y} h'$ после подстановки в (4), получаем

$$\delta I = \int \left[\left(\frac{2y'}{y} h\right)' - h \frac{d}{dt}\left(\frac{2y'}{y}\right) - \left(\frac{y'}{y}\right)^2 h \right] dx. \qquad (6)$$

Для дальнейших выводов примем, что функция плотности вероятностей имеет вид

$$y = f_t(x) = f(x - t), \qquad (7)$$

где t — определяемый параметр сдвига.

В таком случае в (2) y' формально можно считать производной по переменной x. От этого не изменятся выражение (2) и все последующие выкладки. Тогда в (6) можно допустить дифференцирование по x. Исходя из этого, на основании (6), запишем

$$\delta I = \frac{2y'}{y} h \Big|_{x_0}^{x_1} - \int_{x_0}^{x_1} \left[h \frac{d}{dx}\left(\frac{2y'}{y}\right) + \left(\frac{y'}{y}\right)^2 h \right] dx. \qquad (8)$$

При этом в (8) положим $x_0 \to -\infty$, $x_1 \to \infty$, и, в соответствии с этим,

$$\begin{gathered} h(x_1) = h(x_0) = 0 \\ x_1 \to \infty \quad x_0 \to -\infty \end{gathered} \qquad (9)$$

Тогда с учетом (9)

$$\delta I = -\int_{x_0}^{x_1} h(x) \left[\frac{d}{dx}\left(\frac{2y'}{y}\right) + \left(\frac{y'}{y}\right)^2 \right] dx. \qquad (10)$$

Если на функцию y налагаются определенные условия типа

$$\int_{x_0}^{x_1} g(y)\,dx = c, \tag{11}$$

то для поиска экстремума функционала составляется функционал Лагранжа

$$I(y, \lambda) = I + \lambda \int_{x_0}^{x_1} g(y)\,dx, \tag{12}$$

где λ — множитель Лагранжа.

Дифференцированием (12) по Фреше, приравнивая производную подынтегрального выражения к нулю, запишем

$$-\frac{d}{dx}\left(\frac{2\,y'}{y}\right) - \left(\frac{y'}{y}\right)^2 + \lambda g_y{}' = 0. \tag{13}$$

После дифференцирования в (13) по x перепишем его в виде

$$\left(\frac{y'}{y}\right)^2 - 2\,\frac{y''}{y} + \lambda\,g_y{}' = 0. \tag{14}$$

Выражение (14) соответствует дифференциальному уравнению Эйлера—Лагранжа, из решения которого можно найти функцию y.

Рассмотрим решение (14) при следующих условиях:

$$\int y\,dx = 1; \tag{15}$$

$$\int x^2 y\,dx = 1. \tag{16}$$

Суть условия (15) ясна, а суть (16) заключается в том, что дисперсия случайной нормированной величины x известна и равна единице, иными словами, в соответствии с (16) точность измерений известна. Соответствующее (14) уравнение Лагранжа будет иметь вид

$$\left(\frac{y'}{y}\right)^2 - 2\,\frac{y''}{y} + (\lambda_1 + \lambda_2\,x^2) = 0. \tag{17}$$

Для решения (17) перепишем его так:

$$-4\left(\frac{y'}{2\sqrt{y}}\right)'\frac{1}{\sqrt{y}} + (\lambda_1 + \lambda_2\,x^2) = 0,$$

или

$$-\frac{4(\sqrt{y})''}{\sqrt{y}} + (\lambda_1 + \lambda_2\,x^2) = 0. \tag{18}$$

Полагая

$$u=\sqrt{y}, \tag{19}$$

уравнение (18) перепишем в виде

$$-4''+u(\lambda_1+\lambda_2 x^2)=0. \tag{20}$$

При $\lambda_1=-2$ и $\lambda_2=+1$ решением (20) будет

$$u=c\cdot e^{-\frac{x^2}{4}}, \tag{21}$$

а с учетом (19)

$$y=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}. \tag{22}$$

Решение дифференциального уравнения (20) и последующих с помощью рядов приведено в [8].

Если информация о точности измерений отсутствует, условие (16) не будет иметь места, а дифференциальное уравнение Эйлера—Лагранжа примет вид

$$-4u''+\lambda u=0. \tag{23}$$

Решение (23) будет определяться следующими, налагаемыми на функцию плотности распределения, условиями регулярности: 1) непрерывности на всей числовой оси; 2) непрерывности ее первой и второй производной.

1. Если выполняется лишь условие регулярности 1, то при $\lambda=+1$ решением (23) с учетом (19) будет функция

$$y=\frac{1}{2}e^{-|x|}. \tag{24}$$

2. Если выполняются все условия регулярности и, кроме того, известно, что в центре распределения функция плотности терпит максимум, то решением (23) при $\lambda=-4$ будет

$$y=\frac{2}{\pi}\cos^2 x. \tag{25}$$

3. Если же функция плотности, удовлетворяя условиям регулярности, терпит минимум в центре распределения, то при $\lambda=$ $=+4$ решением уравнения Эйлера—Лагранжа будет функция

$$y=\frac{1}{A}ch^2 x, \tag{26}$$

где A — множитель, определяемый из условия нормировки (15).

Простой подстановкой легко убедиться, что из трех функций (24), (25), (26) наименьшее количество информации Фишера доставляется функцией плотности (24).

Следовательно, минимаксной стратегией обработки измерений при неизвестной достоверно точности измерений является распределение Лапласа (24), а при достоверно известной точности измерений — нормальный закон распределения (22).

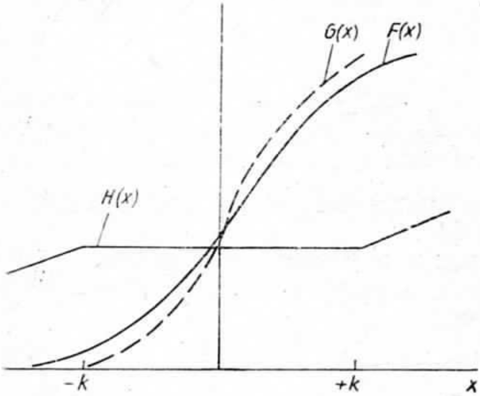

Рис. 1. Функции распределений.

Применение метода максимального правдоподобия к выведенным функциям плотности приводит к оценкам, обладающим по точности их определения минимальным риском.

Приведенные теоретические положения можно применить для построения устойчивого оценивания по отношению к грубым ошибкам. Для этого вначале рассмотрим вероятностную модель совокупности случайных и грубых ошибок. Обозначим функцию распределения и плотность неискаженного грубыми ошибками ряда случайных ошибок соответственно через $G(x)$ и $g(x)$. Распределение грубых ошибок охарактеризуем функцией распределения $H(x)$ и $h(x)$. Долю грубых ошибок в общей совокупности со случайными обозначим через ε. Тогда функция распределения совокупности случайных и грубых ошибок примет вид

$$F(x) = (1-\varepsilon)\, G(x) + \varepsilon H(x), \tag{27}$$

графически это показано на рис. 1.

На рис. 1 k — параметр, разделяющий грубые ошибки от случайных.

Функция плотности, соответствующая (27), будет иметь вид

$$f(x) = \begin{cases} (1-\varepsilon)\, gx, & |x| \leqslant k \\ \varepsilon h(x), & |x| > k. \end{cases} \tag{28}$$

Выражения (27) и (28) доказаны и проиллюстрированы на числовом примере в [9]. В (27) и (28) x — нормированная величина. График, соответствующий (28), показан на рис. 2.

В общем случае модель совокупности случайных и грубых ошибок может определяться выражениями (27) и (28). Однако для получения явного вида функций $f(x)$ и $h(x)$ необходимо знать условия устойчивости, налагаемые на них.

Во-первых, исходя из (28), необходимо, чтобы

$$(1-\varepsilon)\, g(k) = \varepsilon h(k). \tag{29}$$

Во-вторых, необходимо, чтобы функция

$$\psi = [\log f(x)]' \tag{30}$$

была непрерывной. Если же эта функция будет иметь в точке разрыв, то при решении уравнения правдоподобия

$$\int \psi F(dx) = 0 \tag{31}$$

бесконечно малые изменения x могут привести к значительным изменениям значений искомых параметров в зависимости от величины скачка функции ψ. В таких случаях возможна неопределенность в решении. Следствием условия непрерывности является равенство

$$(\log((1-\varepsilon)\, g(x)))' = (\log(\varepsilon h(x)))' \tag{32}$$

при $x=k$.

Необходимо также, чтобы функция ψ была также и ограниченной. При $x \to \infty$ функция ψ будет стремиться к определенной величине и решения (31) будут устойчивыми к большим изменениям x.

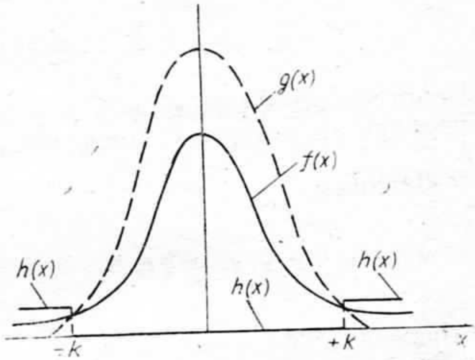

Рис. 2. Плотности распределений.

Таким образом, для получения устойчивого алгоритма уравнивания необходимо, чтобы модель совокупности случайных и грубых ошибок удовлетворяла условиям непрерывности, устойчивости и ограниченности.

В статье поставлена задача определить $h(x)$ и найти зависимость между ε и k при известных $g(x)$, ε.

Для определения $h(x)$ будем исходить из того, что дисперсия грубых ошибок, распределенных на интервалах $[-\infty; -k]$, $[k; \infty]$, не определена. Тогда, исходя из минимаксной стратегии Хьюбера для вывода функции плотности, будем решать дифференциальное уравнение (23).

Отметим, что решением (23) может быть более общее выражение

$$h = ce^{-\lambda_1 |x|}, \tag{33}$$

где c — постоянная, а $\lambda_1 = \sqrt{\lambda}$.

Для определения постоянной c в соответствии с (29) запишем

$$(1-\varepsilon)\,g(k)=\varepsilon\, c e^{-\lambda_1 k}. \tag{34}$$

Тогда

$$c=\frac{1-\varepsilon}{\varepsilon}\,g(k)\,e^{\lambda_1 k}. \tag{35}$$

Для определения λ_1 воспользуемся условием (32). На его основании

$$g'/g=h'/h. \tag{36}$$

Учитывая симметричность функций плотности распределения g и h, после подстановки (33) и (36) находим

$$\lambda_1=-\frac{g'(k)}{g(k)}. \tag{37}$$

Подставляя (37) и (35) в (33), окончательно получаем

$$h=\frac{1-\varepsilon}{\varepsilon}\,g(k)\,e^{\frac{g'(k)}{g(k)}(|x|-k)}. \tag{38}$$

Понятно, что функции g и h должны удовлетворять условию нормировки, которое устанавливает связь между значениями ε и k.

Запишем

$$(1-\varepsilon)\int\limits_{-k}^{+k} g(x)\,dx+2(1-\varepsilon)\,g(k)\,e^{-\frac{g'(k)}{g(k)}k}\int\limits_{k}^{\infty} e^{\frac{g'(k)}{g(k)}|x|}\,dx=1. \tag{39}$$

Выражение (39) является общим для любой симметричной плотности $g(x)$.

Если принять

$$g(x)=\frac{1}{\sqrt{2\pi}}\,e^{-\frac{x^2}{2}}, \tag{40}$$

то, поскольку

$$g'(k)/g(k)=-k, \tag{41}$$

выражение (39) примет вид

$$2(1-\varepsilon)\,\Phi(k)+2(1-\varepsilon)\,g(k)\,e^{k^2}\int\limits_{k}^{\infty} e^{-k|x|}\,dx=1, \tag{42}$$

где $\Phi(k)$ — функция Лапласа.
В (42)

$$\int_k^{\infty} e^{-kx}\,dx = \frac{1}{k}e^{-k^2}, \tag{43}$$

тогда его выражение примет вид

$$2\Phi(k) + 2g(k)\frac{1}{k} = \frac{1}{1-\varepsilon}. \tag{44}$$

Если в (39) принять

$$g(x) = \frac{1}{2}e^{-|x|}, \tag{45}$$

тогда

$$\frac{g'(k)}{g(k)} = -\frac{|k|}{k}, \tag{46}$$

а

$$\int_k^{\infty} e^{-\frac{|k|}{k}|x|}\,dx = e^{-k}. \tag{47}$$

Подставляя (47), (46) в (39), после интегрирования его выражение принимает вид

$$1 = \frac{1}{1-\varepsilon}, \tag{48}$$

которое может быть справедливым лишь при $\varepsilon = 0$. Тогда из (28) следует:

$$f(x) = g(x). \tag{49}$$

Равенство (49) свидетельствует о том, что в случае распределения случайных ошибок по закону Лапласа уравнивание всей совокупности измерений, включающей и грубые ошибки, нужно производить по методу наименьших модулей.

Вернемся теперь к случаю нормального распределения случайных ошибок с функцией плотности (40). Следуя в этом случае выражению (38) на основании (28), запишем

$$f(x) = \begin{cases} \dfrac{1-\varepsilon}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, & |x| \le k \\ \dfrac{1-\varepsilon}{\sqrt{2\pi}} e^{\left(\frac{k^2}{2} - k|x|\right)}, & |x| > k. \end{cases} \tag{50}$$

Аналогичное выражение выведено другим путем в [10]. Здесь же отметим, что выражение, соответствующее условию нормировки (39), в [2] записано с опечатками. Зависимость значений k от ε, полученная на основании (44), приведена ниже:

k	0,1	0,2	0,3	0,4	0,5	0,07	0,7	0,8	0,9	0,10
ε	0,88	0,75	0,64	0,54	0,44	0,36	0,29	0,23	0,19	0,14
k	1,5	2,0	2,5	3,0	—	—	—	—	—	—
ε	0,04	0,007	0,002	0,0001	—	—	—	—	—	—

Как следует из (28), (50), функция плотности распределения совокупности случайных и грубых ошибок имеет сложный вид. Это в свою очередь усложняет функцию правдоподобия. Например, исходя из (50), ее вид в случае ненормированной случайной величины будет

$$\psi = \begin{cases} \dfrac{x-t}{\sigma^2}, & |x-t| \leqslant k\sigma \\ \dfrac{1}{\sigma} k \operatorname{sign}(x-t), & |x-t| > k\sigma. \end{cases} \tag{51}$$

Тогда решение уравнения правдоподобия (31) возможно здесь лишь итерационным способом. При этом из разложения (31) в ряд Тейлора находим

$$t - t_0 = -\frac{\int \psi_0 F(dx)}{\int \psi_0' F(dx)}. \tag{52}$$

Выражение (52) неудобно в случаях, когда $\Psi' = 0$, поэтому, учитывая, что [1]

$$\int \left(\frac{f'}{f}\right)' f dx = -\int \left(\frac{f'}{f}\right)^2 f dx, \tag{53}$$

перепишем его в виде

$$t = t_0 + \frac{\int \psi_0 F(dx)}{\int (\psi_0)^2 F(dx)}. \tag{54}$$

Отметим, что здесь, как и ранее, принято

$$F(dx) = f dx. \tag{55}$$

Для удобства вычислений выражение (53) запишем так:

$$t_n = t_{n-1} + \frac{\int \psi_{n-1} p}{\int (\psi_{n-1})^2 p}, \tag{56}$$

где p — относительная частота.

В случае (51) выражение (56) примет вид

$$t_n = t_{n-1} + \frac{\left(\sum_1^M \frac{(x - t_{n-1})}{\sigma^2} + \frac{k}{\sigma} \sum_1^N \operatorname{sign}(x - t_{n-1}) \right) \frac{1}{m}}{\left(\sum_1^M \frac{(x - t_{n-1})^2}{\sigma^4} + N \frac{k^2}{\sigma^2} \right) \frac{1}{m}},$$

где M — число значений $(x - t_{n-1})$, не превышающих $k\sigma$; N — число таких значений, превышающих $k\sigma$.

В качестве примера рассмотрим результаты моделирования линейных измерений. Ниже приведены значения отклонений результатов измерений от устойчивого значения длины:

x, м	102,361	369	370	350	380	460	520
$x-t$, мм	−9	−1	0	−20	+10	+90	+140

Истинное значение длины принималось равным 102,370 м. Стандарт измерений 10 мм. Грубыми ошибками искажены шестое и седьмое измерения, т. е. $\varepsilon = 0{,}29$. Распределение случайной части ошибок — нормальное. Необходимо определить устойчивое значение измеряемой длины линии.

В качестве нулевого приближения длины примем медиану вариационного ряда измерений, т. е. $t_0 = 102{,}370$ м.

В [9] показано, что медиана является оценкой метода наименьших модулей в одномерном случае.

Тогда при $k\sigma = 0{,}7 \cdot 10 = 7$ мм найдем

$$t_1 = 370 + \frac{(0-1) + 7 \cdot (-1 - 1 + 1 + 1 + 1)}{\frac{1}{100} + 5 \cdot 0{,}4} = 373.$$

Во втором приближении снова $t_2 = 373$. На этом заканчиваем вычисления.

Если в качестве нулевого приближения брать крайние значения, например 510 или 350, то в результате получаем точно такой же результат, но число приближений в первом случае 19, а во втором — три.

Таким образом, на основе вариационного подхода вопрос вывода функций плотности для устойчивого оценивания параметра положения по отношению к грубым ошибкам решается минимизацией количества информации Фишера.

1. *Боровков А. Я.* Математическая статистика. М., 1984. 2. *Вучков И. Н., Бояджиева Л. Н., Солаков Е. Б.* Прикладной линейный регрессионный анализ. М., 1987. 3. *Гельфанд И. М., Фомин С. В.* Вариационное исчисление. М., 1961. 4. *Колмогоров А. Н.* Теория вероятностей и математическая статистика. М., 1986. 5. *Маркузе Ю. И., Юркина М. И.* Тез. док. VII Международного симпозиума по геодезическим вычислениям // Геодезия и картография. 1985. № 10. С. 58—61. 6. *Никифоров Б. И.* Замечания по статье М. Л. Маргулева //

Геодезия и картография. 1986. № 6. С. 18—21. 7. *Сухов А. Н.* Дифференциальная функция погрешностей измерений // Геодезия, картография и аэрофотосъемка. 1990. Вып. 52. С. 65—71. 8. *Ярмоленко А. С.* Минимаксное оценивание и устойчивость при математической обработке геодезических измерений. — Горки, 1988. — 52 с. — Рукопись деп. в ВНИТИ, № 5293—В88. 9. *Ярмоленко А. С.* Алгоритм устойчивого способа уравнивания геодезических измерений и минимаксное оценивание. Горки. 1989. — 180 с. — Рукопись деп. в ВИНИТИ, № 4455—В89. 10. *Huber P. I.* Robust statistics. N.-Y. Willey, 1981.

Статья поступила в редколлегию 03. 09. 91
