А. Л. ПЕТРОСЯН

ОБ ОПРЕДЕЛЕНИИ ПЕРИОДА ИЗ ЗАПИСИ

Существует ряд задач инженерной сейсмологии, в которых важно уметь возможно точнее измерять период колебания. Назовем здесь хотя бы определение резонансных свойств грунтов по преобладающим периодам микроколебаний и местных землетрясений (см. [2—4]), как и вообще анализ записей колебаний систем с дискретными характерными периодами.

Искусственный пример записи колебания с приблизительно постоянным периодом показан на рис. 1. Обычно измеряют время появления первой и последней вершин, t_1 и t_n, и затем вычисляют так называемый «средний» период по формуле

$$\bar{T}=(t_n-t_1)/\frac{n-1}{2}. \tag{1}$$

Легко видеть, что эта формула могла бы быть применена к любым другим двум вершинам на данной записи, хотя результат имел бы бо́льшую ошибку. Это означает, что используя только две крайние вершины, мы теряем значительную часть информации, заключенной в записи. Нужно, чтобы в вычислении периода участвовали все n измерений времен прихода вершин. Выведем соответствующие формулы.

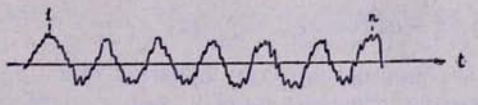

Рис. 1. Пример записи колебания с постоянным периодом искаженного шумом

Предположение о постоянном периоде записи требует, чтобы «истинные» времена появления вершин (т. е. времена, которые можно было бы измерить на записи, полностью освобожденной от шума) $\tau_1,\ldots,\tau_n$ удовлетворяли соотношению

$$\tau_i=\tau_1+\frac{T}{2}(i-1). \tag{2}$$

Положим, что результат измерения величины τ_i представляет собой наблюдение случайной величины ζ_i, распределенной нормально с ожиданием τ_i и дисперсией σ^2. Тогда, максимизируя совместную плотность вероятности величин $\zeta_1,\ldots,\zeta_n$ при $\zeta_i=t_i$,

$$(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[t_i-\tau_1-\frac{T}{2}(i-1)\right]^2\right.,$$

получаем оценку максимального правдоподобия для T:

$$\hat{T}=\frac{24}{n(n^2-1)}\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)t_i. \tag{3}$$

Её дисперсия

$$D^2\{\hat{T}\}=\frac{48}{n(n^2-1)}\sigma^2,$$

а дисперсия обычной оценки по формуле (1)

$$D^2\{\bar{T}\}=\frac{8}{(n-1)^2}\sigma^2.$$

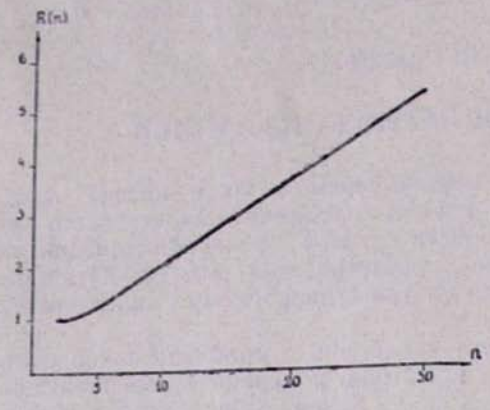

Рис. 2. Отношение дисперсий, $R(n)=\frac{D^2\{\bar{T}\}}{D^2\{\hat{T}\}}$ обычной оценки периода $\bar{T}$ и предложенной $\hat{T}$ как функция числа наблюдений n

Их отношение

$$R(n)=\frac{D^2\{\bar{T}\}}{D^2\{\hat{T}\}}=\frac{n(n+1)}{6(n-1)},$$

как функция числа наблюдений n изображено на рис. 2, из которого видно, что дисперсия предлагаемой оценки при любом n, большем трёх, меньше дисперсии обычной оценки, а при $n=3$ равна ей.

Заметим, что наша задача в постановке (2) является задачей о линейной регрессии на неслучайную независимую переменную i. Напишем соотношение (2) в виде

$$\tau_i=\alpha+\beta(x_i-\bar{x}),$$

где β—искомый период, $x_i=\frac{i}{2}$ и значит $\bar{x}=\frac{n+1}{4}$, а α—некоторая постоянная, которая определится позже. Теперь имеем ([5], § 37.2) оценки максимального правдоподобия

$$\hat{\alpha}=\bar{t}$$

$$\hat{\beta}=\hat{T} \qquad \text{(см. формулу (3))}$$

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}\left[t_i-\hat{\alpha}-\hat{\beta}\left(\frac{i}{2}-\frac{n+1}{4}\right)\right]^2.$$

Здесь $\hat{\alpha}$ и $\hat{\beta}$ независимы, а $n\hat{\sigma}^2/\sigma^2$ распределена как χ^2 с $n-2$ степенями свободы. Следовательно, величина

$$\tau=\sqrt{\frac{(n-2)(n^2-1)}{48}}\,\frac{\hat{\beta}-\beta}{\hat{\sigma}}=\sqrt{\frac{(n-2)(n^2-1)}{48}}\,\frac{\hat{T}-T}{\hat{\sigma}}=\tau(\hat{T}-T) \qquad (4)$$

имеет распределение Стьюдента с $n-2$ степенями свободы. С помощью (4) можно проверять гипотезы относительно периода T, строить доверительные интервалы для него и т. д. Например, $p\%$-й доверительный интервал для T определяется неравенством

$$\hat{T}-\tau^{-1}t_p<T<\hat{T}+\tau^{-1}t_p,$$

где t_p отыскивается по таблицам распределения Стьюдента для n—2 степеней свободы.

В качестве примера рассмотрим определение периода микросейм по записи, состоящей из десяти вершин (n=10). Результаты времен появления вершин (в секундах) следующие: 16,6; 18,6; 20,8; 23,6; 26,4; 30,4; 32,8; 35,6; 38,4; 41,2. Отсюда по вышеприведенным формулам получаем $\hat{T}=5{,}64$, $\hat{\sigma}=0.467$. Левый конец 95%-го доверительного интервала

$$T_1=\hat{T}-\sqrt{\frac{48}{(n-2)(n^2-1)}}\hat{\sigma}t_p=5.37,$$

а правый

$$T_2=\hat{T}+\sqrt{\frac{48}{(n-2)(n^2-1)}}\hat{\sigma}t_p=5.91,$$

т. е. $5.37<T<5.91$ на уровне доверия 95%.

До сих пор мы рассматривали оценку периода из одиночной записи. Однако микросейсмы и микроколебания появляются на сейсмограмме в виде большого числа правильных групп колебаний, в которых можно измерять времена появления вершин. Эти группы чередуются с участками нерегулярных колебаний меньшей амплитуды, где измерения затруднены. В каждой такой правильной группе можно получить оценку периода вышеуказанным способом, и тогда возникает проблема объединения этих оценок.

Эта задача известна как задача об объединении неравноточных наблюдений. Ей уделяется немало места в руководствах по обработке астрономических [8], физических [9] и геодезических наблюдений [7]. Математически эта задача обычно формулируется так: имеется k выборок вида $x_{i1},\ldots, x_{ini}$, извлеченных из нормальной совокупности $N(\mu, \sigma_i^2)$, где n_i и σ_i^2 могут меняться произвольным образом от выборки к выборке. Требуется найти оценку параметра μ.

Задача о периоде имеет два отличия. Во-первых, в сумме квадратов ns_i^2 внутри каждой выборки будет уже не n_i-1, а n_i-2 степеней свободы. Во-вторых, и это существенно, в случае равных дисперсий $\sigma_1^2=\ldots=\sigma_k^2$ нельзя просто взять среднеарифметическое всех наблюдений, что, как известно, является наилучшей оценкой, так как каждая выборка имеет свое начало отсчета τ_1 (или, что совершенно равносильно, $\bar{\tau}$).

Трудность рассматриваемой задачи заключается в том, что помимо интересующего нас параметра μ здесь с каждой новой выборкой прибавляется ненужный нам (и неизвестный) параметр σ_i^2. Впервые такого рода задачи и связанные с ними трудности были рассмотрены в работе [13], где были введены термины структурный и мешающий параметр.

Первое по времени своего появления решение поставленной задачи заключается в следующем. Вычислив из каждой выборки величины

$$\hat{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij};\quad s_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\hat{X}_i)^2, \tag{5a}$$

составим веса

$$w_i=s_i^{-2}.$$

Оценка $\bar{\mu}_w$ получается теперь по формуле

$$\bar{\mu}_w = \sum_{i=1}^{k} W_i \hat{X}_i \Big/ \sum_{i=1}^{k} W_i. \tag{5}$$

Эта оценка известна как взвешенное среднее ([9], гл. 11).

Принцип наибольшего правдоподобия в теории оценивания является очень общим методом нахождения оценок. Он применим также и в этом случае. Совместная плотность вероятности для отдельной выборки пропорциональна

$$\sigma_i^{-n_i} \exp\left\{ -\frac{n_i(\hat{X}_i-\mu)^2+(n_i-1)s_i^2}{2\sigma_i^2} \right\}$$

(см. [6], § 26). Остальная часть совместной плотности нас не интересует, так как не зависит от μ и σ_i^2. Функция правдоподобия для k выборок есть

$$L = \left(\prod_{i=1}^{k} \sigma_i^{-n_i} \right) \exp\left\{ \sum_{i=1}^{k} \left(-\frac{n_i(\hat{X}_i-\mu)^2+(n_i-1)s_i^2}{2\sigma_i^2} \right) \right\}.$$

Дифференцируя $\ln L$ по μ и σ_i^2 и приравнивая полученные выражения нулю, получаем

$$\left.\begin{aligned} \sum_{i=1}^{k} \frac{n_i(\hat{X}_i-\mu)}{\sigma_i^2} &= 0 \\ \sigma_i^2 &= \frac{n_i-1}{n_i} s_i^2+(\hat{X}_i-\mu)^2 \end{aligned}\right\} \tag{6}$$

Подстановка σ_i^2 под знак суммы дает уравнение для определения искомой оценки

$$\sum_{i=1}^{k} \frac{n_i^2(\hat{X}_i-\mu)}{(n_i-1)s_i^2+n_i(\hat{X}_i-\mu)^2} = 0. \tag{7}$$

Нейман и Скотт [7] вывели другую оценку $\hat{\mu}_{NS}$, которая дается корнем уравнения

$$\sum_{i=1}^{k} \frac{n_i(n_i-2)(\bar{X}_i-\mu)}{(n_i-1)s_i^2+n_i(\hat{X}_i-\mu)^2} = 0, \tag{8}$$

и доказали, что ее асимптотическая дисперсия меньше таковой для $\hat{\mu}$. Заметим, однако, что эта оценка обладает одной примечательной особенностью [10]. Дело в том, что при $n_i=2$ соответствующие члены в сумме (8) равны нулю. Это означает, что выборки, состоящие из двух измерений, не дают вклада в эту оценку, так что, например, если все $n_i=2$, то $\hat{\mu}_{NS}$ просто неопределима. Это весьма нежелательное свойство, так как выборки с $n_i=2$ обладают оценкой разброса внутри выборки (s_i^2) и поэтому могут участвовать в образовании взвешенного среднего и оценки наибольшего правдоподобия (7).

Работа [11] посвящена различным способам извлечения информации из функции правдоподобия для случая, когда имеется большое число мешающих параметров, где, в частности, рассматривается и настоящая задача. Оценка $\hat{\mu}_c$ является корнем уравнения

$$\sum_{i=1}^{k} \frac{n_i(n_i-1)(\hat{X}_i-\mu)}{(n_i-1)s_i^2+n_i(\hat{X}_i-\mu)^2} = 0. \tag{9}$$

Уравнения (7—9) отличаются друг от друга лишь множителем в числителе, который принимает значения соответственно n_i, n_i-2, и n_i-1.

Наконец, приведем предлагаемую нами оценку. Будем рассуждать следующим образом. Возьмем некоторое гипотетическое значение μ_0 параметра μ и образуем k статистик Стьюдента

$$z_i = \frac{\hat{X}_i-\mu_0}{S_i} \quad (i=1,2,\ldots,k), \tag{10}$$

где $S_i^2 = \frac{S_i^2}{n_i}$ —выборочная дисперсия величины $\hat{X}_i$. Преобразование, введенное Р. А. Фишером [10] и подробно разобранное в статье Е. Пирсона [14]

$$p_i = \int_{-\infty}^{z_i} f_{n_i-1}(x)\,dx, \tag{11}$$

переводит случайную величину z_i, имеющую распределение Стьюдента с n_i-1 степенями свободы, в случайную величину p_i, имеющую равномерное распределение на интервале (0,1). В формуле (11) $f_m(x)$ есть плотность распределения Стьюдента с m степенями свободы. Если истинное значение μ больше μ_0, то ожидание

$$E\{\hat{X}_i\} > \mu_0$$

и вероятность положительных значений z_i будет больше, чем отрицательных, так что $p_1,\ldots,p_k$ сгустятся к правому краю интервала (0,1). Тогда для проверки гипотезы $H_0: \mu=\mu_0$ против $H_1: \mu>\mu_0$ нужно использовать статистику [14]

$$V_+ = -2\sum_{i=1}^{k} \ln(1-p_i), \tag{12а}$$

а против $H_2: \mu<\mu_0$—статистику

$$V_- = -2\sum_{i=1}^{k} \ln p_i. \tag{12б}$$

Обе эти статистики распределены как χ^2 с $2k$ степенями свободы.

V_- и V_+, вычисляемые из наблюдений, дают соответствующие им уровни значимости (из таблиц распределения χ^2). Будем рассматривать эти уровни значимости как численное выражение свидетельства, доставляемого выборками, против соответственно гипотез $H_2: \mu<\mu_0$ и $H_1: \mu>\mu_0$. Предлагаемая нами оценка $\bar{\mu}$ является корнем уравнения

$$V_- = V_+, \tag{13}$$

т. е. $\tilde{\mu}$ представляет собой такое значение параметра μ, которое отвергается статистиками V_- и V_+ на одинаковом уровне значимости. Заметим, что если бы мы пытались построить критерий не против односторонних альтернатив H_1 и H_2, а против H_3: $\mu \neq \mu_0$, то такой метод оценки был бы невозможен.

Исследуем теперь сравнительные достоинства этих различных оценок. Аналитическое исследование этого вопроса представляет значительные трудности, когда речь идет о малых выборках. Например, в уже цитированной большой статье [13] доказан лишь тот довольно слабый результат, что асимптотическая (при $k \to \infty$) дисперсия $\hat{\mu}_{NS}$ из (8) меньше дисперсии для $\hat{\mu}$ (7). Для практического применения важно знать, насколько меньше, чтобы иметь возможность судить о целесообразности замены одной оценки другой. Естественно поэтому использовать метод Монте Карло (имитация повторных случайных выборок, обычно на ЭВМ с помощью псевдослучайных чисел).

Дополнительной и характерной трудностью для данного случая в выяснении сравнительных достоинств оценок является неопределенность в выборе $n_1, \ldots, n_k$ и $\sigma_1^2, \ldots, \sigma_k^2$. Мы ограничимся имитацией одной экспериментальной ситуации, а именно, представим себе, что наши выборки—это результаты измерений одной и той же величины несколькими приборами различной точности. Легко понять, что измерения высокоточным прибором должны стоить дороже, чем более грубым, так что мы имеем возможность провести лишь несколько измерений первым, но зато сравнительно много вторым, компенсируя этим низкую точность самих измерений.

Мы взяли $k=10$—десять выборок, числа наблюдений в них, $n_1, n_2, \ldots, n_{10}$ равны 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, а стандартные отклонения $\sigma_1, \sigma_2, \ldots, \sigma_{10}$ возрастают с номером выборки:

Случай 2:1;1,1; 1,2; 1,3;..., 1,9.

Случай 2:1,1,1; 1,2; 1,3;..., 1,9.

Размеры выборок и стандартные отклонения здесь подобраны так, что с уменьшением точности измерений возрастает их количество.

Представляет интерес и исследование противоположной ситуации—чем точнее измерения, тем больше их сделано:

Случай 3: 1,9; 1,8; 1,7;..., 1,1; 1;

Случай 4: 2,4; 2,2; 2,..., 0,8; 0,6.

Процесс счета состоял в следующем. Для $n_1, \ldots, n_{10}$, равных 2, 3,..., 10, 15 и $\sigma_1, \ldots, \sigma_{10}$, выписанных выше (случаи 1—4) с помощью подпрограммы псевдослучайных чисел, извлекались десять выборок $i=1,2,\ldots,10$ нормально распределенных чисел с общим ожиданием μ и стандартным отклонением σ_i. Из каждой выборки вычислялись $\bar{X}_i$ и s_i^2 (см. (5а)). Далее вычислялись взвешенное среднее (5) и $\hat{\mu}$, $\hat{\mu}_{NS}$, $\hat{\mu}_c$, которые являются корнями по μ соответственно уравнений (7), (8), (9). Наша оценка $\tilde{\mu}$ является корнем уравнения (13), где V_+ и V_- даются формулами (12а, б), а p_i находится из таблиц распределения Стьюдента [1] с аргументом (10) для прямых измерений ожидания μ и с аргументом (4) для оценок периода по отдельным записям.

Весь этот процесс извлечения десяти выборок и вычисления оценок повторялся в среднем $N=200$ раз для каждого из четырех случаев. Для

каждой из оценок вычислялось среднеквадратичное отклонение от известного нам истинного значения ожидания μ. В случае нашей оценки $\tilde{\mu}$ это будет

$$\Sigma_{\tilde{\mu}} = \frac{1}{N}\sum_{n=1}^{N}(\tilde{\mu}_n - \mu)^2$$

В табл. 1 приведены результаты счета.

Таблица 1

Номер случая	$\sigma_K - \sigma_1$	(5)	(7)	(8)	(9)
1	2,8	1,43	1,61	1,60	1,58
2	0,9	3,06	2,02	1,30	1,48
3	—0,9	5,62	1,47	1,04	1,42
4	—1,8	5,95	3,38	0,69	1,51

$\sigma_k - \sigma_1$—суммарное увеличение стандартного отклонения от первой выборки к последней. В колонках, обозначенных номерами 5, 7, 8, 9 (номера уравнений, по которым вычисляются оценки), стоят отношения среднеквадратичных отклонений этих оценок (соответственно $\bar{\mu}_W$, $\hat{\mu}$, $\hat{\mu}_{NS}$, $\hat{\mu}_{c<}$) к отклонению предлагаемой оценки $\tilde{\mu}$, т. е. к $\Sigma_{\tilde{\mu}}$. Таким образом, чем больше это отношение, тем хуже соответствующая оценка по сравнению с предлагаемой $\tilde{\mu}$. Цифры в таблице говорят сами за себя. Лишь в случае 4 оценка $\hat{\mu}_{NS}$ лучше нашей, но причина этого ясна: множитель (n_1-2) дает относительно меньшие веса малым выборкам, т. е. как раз тем, которые в данном случае имеют большие σ_1 и портят оценку. Конечно, на практике, если есть много хороших наблюдений и мало плохих, последние просто отбрасывают.

Ордена Ленина Институт физики Земли АН СССР

ЛИТЕРАТУРА

1. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М., 1965.
2. Ершов И. А. Сопоставление инструментальных данных о скоростях распространения волн в грунте, амплитудах и периодах для сейсмического микрорайонирования. Сб. «Сейсмическое микрорайонирование» (Вопросы инженерной сейсмологии, вып. 10), М., «Наука», 1965.
3. Ершов И. А. Об использовании микроколебаний для определения частотных особенностей грунтов. Сб. «Сейсмическое движение грунта» (Вопросы инженерной сейсмологии, вып. 13), М., «Наука», 1970.
4. Ершов И. А., Медведев С. В., Федотов С. А., Штейнберг В. В. Сейсмическое микрорайонирование Петропавловска-Камчатского. Сб. «Сейсмическое микрорайонирование» (Вопросы инженерной сейсмологии, вып. 10), М., «Наука», 1965.
5. Крамер Г. Математические методы статистики. Перев. с англ., изд. второе, М., «Мир», 1975.
6. Романовский В. И. Математическая статистика. Кн. 2, Ташкент, 1963.
7. Смирнов Н. В., Белугин Д. А. Теория вероятностей и математическая статистика в приложении к геодезии. М., «Недра», 1969.
8. Щиголев Б. М. Математическая обработка наблюдений. М., Физматгиз, 1962.

9. Яношя Л. Теория и практика обработки
«Мир», 1968.
10. Fisher R. A. Statistical methods for research workers. 4th ed., London, Oliver & Boyd. 1931
11. James G. S. The Behrens-Fisher distribution and weighted means. J. Roy. Statist, Soc. B. 21, 2. 1959.
12. Kalbfleisch J. D. and Sprott D. A. Application of likelihood methods to models involving large numbers of parameters (with discussion). J. Roy. Statist. Soc. B. 32. 2. 1970.
13. Neyman J. and Scott. Elizabeth L. Consistent estimates based on partially consistent observations. Econometrica 16. 1. 1948.
14. Pearson E. S. The probability integral transformation for testing goodness of fit and combining independent tests of significance. Biometrika, Parts 1 and 2, 1938.