АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ

№ 6 1970

УДК 519.24

В. С. КИРИЧУК, Б. Н. ЛУЦЕНКО

(Новосибирск)

ИСКЛЮЧЕНИЕ НЕДОСТОВЕРНЫХ ДАННЫХ

Вопрос проверки экспериментальных данных на достоверность привлекает внимание исследователей вот уже более ста лет. В настоящее время четко обозначились две тенденции. Первая — использование аномальных данных в обработке с соответствующими весами. Начало этому направлению положил немецкий астроном Бессель в 1838 г. Такой подход требует знания параметров распределения аномальных измерений (выбросов). Второй подход сводится к выявлению и изъятию из обработки недостоверных данных. Сфера действия его несколько шире.

В большинстве работ анализируется простая модель — повторная выборка, по которой необходимо оценить математическое ожидание в виде константы. Далеко не всегда конкретную задачу удается привести к этой модели.

Для более детального ознакомления с кругом идей, используемых в области анализа аномальных данных, можно обратиться к последней из обзорных статей [1]. Она содержит дальнейшие ссылки на работы 15 предшествующих лет. Из более поздних публикаций упомянем лишь некоторые, наиболее близкие к интересующей нас задаче.

В [2] используется изящный прием исправленных разностей. Суть его вкратце в следующем. Формируется вспомогательный нормальный вектор и суммируется с вырожденным вектором разности между измеренными и обработанными значениями. Причем вспомогательный вектор выбирается с таким расчетом, чтобы результирующий вектор имел невырожденные компоненты с нулевым средним и одной и той же дисперсией. Последующий анализ производится уже с этим вектором исправленных разностей. Это позволяет обойтись одномерным интегралом вместо n-мерного при обычном подходе [3]. Для использования этого алгоритма необходимо знать дисперсию генеральной совокупности и параметры закона распределения выбросов. Несмотря на громоздкость предварительных вычислений в случае общей линейной модели сам алгоритм весьма прост. Однако, если вид модели сигнала меняется в ходе обработки, использование метода затруднительно.

В [4] для обработки выборки, содержащей выбросы, используется мощный аппарат условных распределений. Алгоритм оптимален, некритичен к выбору параметров распределения выбросов, но при большой выборке и в случае общей линейной модели даже после упрощений требует фантастического объема вычислений.

Перспективным, с нашей точки зрения, является предложенное в [5] использование в проблеме выбросов аппарата теории статистических решений. Основным достоинством алгоритма является исключительная простота. Априорное распределение выбросов предполагается известным. К сожалению, автор не приводит никаких соображений по выбору исходного базисного участка (выборки, не содержащей выбросов). Оптимальность метода очевидна лишь при наличии одного выброса. Предполагаемая автором сходимость и оптимальность последовательной процедуры требует более четкого обоснования. На практике весьма редко бывает заранее известно распределение выбросов, кроме того, немаловажную роль играет также и простота алгоритма. В предлагаемой работе предпринимается попытка построить критерий, не использующий априорного распределения выбросов и требующий несложных вычислений. Пусть имеется комплекс n измерительных приборов, поставляющих информацию о некоем процессе. Сам процесс описывается совокупностью параметров $A^T = \{\alpha_1, \ldots, \alpha_m\}$. Показания каждого из приборов нелинейным образом зависят от A и сопровождаются случайными погрешностями

$$C_i = f_i(\alpha_1, \ldots, \alpha_m) + \xi_i. \tag{1}$$

Кроме того, некоторые из приборов в определенные моменты могут поставлять недоброкачественные данные, содержащие погрешности, намного превосходящие величину, предусматриваемую моделью случайной компоненты ξ_i, или маловероятные, с точки зрения этой модели, так называемые выбросы.

Объединяя все уравнения вида (1) в единое матричное уравнение, получим

$$C = F(\alpha_1, \ldots, \alpha_m) + \Xi, \tag{2}$$

причем предполагается, что

$$\Xi \in N(0, \sigma^2 K),$$

т. е. что вектор Ξ нормален с нулевым вектором средних и известной корреляционной матрицей K. Будем предполагать, кроме того, что величина σ^2 достаточно мала, что позволит в дальнейшем при отыскании оценки вектора A воспользоваться линеаризацией и считать оценки приближенно нормальными и несмещенными.

Оценку вектора A отыскиваем известным способом. Задавшись нулевым приближением A_0 вектора A, линеаризуем в его окрестности соотношение (2)

$$C \cong F(\alpha_{10}, \ldots, \alpha_{m0}) + \left(\frac{\partial F}{\partial A}\right)_{A = A_0} (A - A_0) + \Xi. \tag{3}$$

Здесь $\left(\frac{\partial F}{\partial A}\right)_{A = A_0} = D_0$ — матрица, составленная из частных производных измеряемых величин по искомым параметрам, взятых в точке A_0:

$$d_{ij} = \left(\frac{\partial f_i(\alpha_1, \ldots, \alpha_m)}{\partial \alpha_j}\right)_{A = A_0}.$$

Далее из (3) находим оценку для вектора поправок $\Delta A_0 = A - A_0$, исходя из принципа максимального правдоподобия

$$\Delta \hat{A}_0 = (D_0^T K^{-1} D_0)^{-1} D_0^T K^{-1} [C - F(A_0)]. \tag{4}$$

Уточняем оценку A_0: $A_1 = A_0 + \Delta \hat{A}_0$ — и продолжаем описанную процедуру

до вхождения вектора поправок ΔA_ν в назначенную заранее ε-окрестность.

При постулируемой малости погрешностей можно считать вектор $\hat{A}_\nu$ приближенно нормальным и несмещенным, в результате чего вектор

$$\Delta C_\nu = C - F(\hat{A}_\nu) \cong [I - D_\nu (D_\nu^T K^{-1} D_\nu)^{-1} D_\nu^T K^{-1}] \Xi \qquad (5)$$

приближенно нормален:

$$\Delta C_\nu \in N(0,\ B_{\Delta C}),$$

причем

$$B_{\Delta C} = \sigma^2 [K - D (D^T K^{-1} D)^{-1} D^T] = \sigma^2 K_{\Delta C}.$$

Основываясь на этом, можно выделить наиболее подозрительные (в смысле наличия выбросов) измерения, взвесив компоненты ΔC_i, согласно их дисперсиям:

$$\delta_i = \frac{\Delta C_i}{\sqrt{K_{ii\,\Delta C}}}.$$

(Заметим, что просто декоррелировать компоненты вектора ΔC мы не имеем возможности, поскольку он вырожден.)

Однако выделить подозрительные измерения еще недостаточно. Необходимо иметь основание отбросить или включить их в обработку. Этот этап является наиболее трудным в проблеме выбросов. Различные подходы к его решению и вызвали к жизни столь пестрое разнообразие алгоритмов.

Относительно выбросов будем предполагать, что их появление равновероятно в каждом измерении.

Допустим вначале, что дисперсия генеральной совокупности известна априори. В этих обстоятельствах равномерным наиболее мощным критерием [6] отбраковки является критерий, построенный следующим образом. Вектор ΔC подвергнем линейному преобразованию так, что в новой системе координат вектор ΔC^* будет иметь лишь $n - m$ не равных тождественно нулю равноточных компонент.

Осуществим эту операцию в два приема. Вначале декоррелируем вектор C:

$$C^* = K^{-1/2} C.$$

Здесь $K^{-1/2}$ — символическое обозначение матрицы, определяемой из условия

$$K^{-1/2} K^{-1/2} = K^{-1}.$$

Выражение (5) преобразуется в

$$\Delta C^* \cong [1 - D^* (D^{*T} D^*)^{-1} D^{*T}] \Xi^* = N \Xi^*,$$

где

$$D^* = K^{-1/2} D; \quad \Xi^* = K^{-1/2} \Xi.$$

Здесь N — матрица проектирования на $n - m$-мерное пространство. Перейдем к новой системе координат, $n - m$ первых осей которой лежат

в пространстве, определяемом N:

$$\Delta^* = U^* N \Xi^* = \begin{pmatrix} \Delta_1 \\ \cdot \cdot \cdot \\ \Delta_{n-m} \\ 0 \\ \cdot \cdot \cdot \\ 0 \end{pmatrix}.$$

Здесь U^* — матрица поворота.

Первые $n-m$ строк вектора Δ^* и матрицы U^* содержат всю информацию исходной выборки. Матрица U определяется из условия

$$UN=U,$$

в силу того, что U (матрица, составленная из первых $n-m$ строк U^*) определяет то же пространство, что и N. При практических вычислениях можно выбрать первую строку U состоящей из $m+1$ первых отличных от нуля компонент, вторую из $m+2$ и т. д. Это позволит записать для вычисления элементов U рекурсивную процедуру, основанную на блочном обращении матриц [7].

При отсутствии в результатах измерений выбросов все компоненты Δ_{n-m} независимы и имеют одну и ту же дисперсию

$$\Delta_{n-m} \in N(0, \sigma^2 I).$$

Задавшись доверительной вероятностью P_0, определим уровень отбраковки l из условия

$$P(|\Delta_{\max}| \leqslant l\sigma) = P_0,$$

где l определяем как $\alpha/2$-квантиль нормального распределения

$$\alpha = 1 - P_0^{1/n}.$$

Подозрительное измерение изымается из обработки, если $|\Delta_{\max}| > l\sigma$, и процедура повторяется до тех пор, пока не выполнится условие

$$|\Delta_{\max}| \leqslant l\sigma.$$

К сожалению, при больших объемах выборки критерий требует громоздких вычислений, связанных с нахождением матрицы U.

В этих условиях предлагается другой, менее оптимальный, но зато требующий меньшего объема вычислений критерий.

Подчиним алгоритм отбраковки требованию обеспечения максимальной выборочной плотности в окрестности точечной оценки $\hat{A}$. Рассмотрим ради этого доверительный m-мерный эллипсоид с центром в $\hat{A}$

$$\frac{\frac{1}{\sigma^2}(A-\hat{A})^T K_{\hat{A}}^{-1}(A-\hat{A})}{\frac{1}{\sigma^2(n-m)}\Delta C^T K^{-1}\Delta C} = \varphi. \qquad (6)$$

Числитель и знаменатель (6) статистически независимы и распределены по χ^2 с m и $n-m$ степенями свободы. В обычных условиях φ имеет распределение Фишера. Мы же производим упорядочение выборки и на каждом шаге все более ограничиваем величину

$$S^2 = \frac{\Delta C^T K^{-1} \Delta C}{n-m}$$

Будем полагать, что

$$S^2_{\nu-1} < S^2_{\nu},$$

где S^2_{ν} — оценка дисперсии, полученная на ν-м шаге отбраковки. (Это равносильно предположению правильной отбраковки выбросов.) Соответственно распределение S^2/σ^2 принимается усеченным на уровне $S\nu^2/\sigma^2$. Введем новую переменную $W = (n-m)\sigma^2 S^2$.

Выражение для совместной плотности $\varphi_и$ W приведено в [8]. Нам остается лишь добавить нормирующий множитель $Q(n, \mu)$, обусловленный ограниченностью W:

$$\Psi(W, \varphi) = Q(n, \mu) f(W, \varphi)\, d\varphi dW,$$

где

$$f(W, \varphi) = \frac{\left(\frac{m}{n-m}\right)^{\frac{m}{2}}}{2\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n-m}{2}\right)} \varphi^{\frac{m}{2}-1} \left(\frac{W}{2}\right)^{\frac{n}{2}-1} e^{-\frac{1}{2}\left[1+\frac{m\varphi}{n-m}\right]W}$$

Определим $Q(n, \mu)$ из условия

$$Q(n, \mu) \int_0^{\infty} \int_0^{2\mu} f(W, \varphi)\, d\, W\, d\varphi = 1. \tag{7}$$

Здесь $\mu = \frac{n-m}{2\sigma^2}\, S^2_{\nu}$. Рассмотрим

$$\int_0^{\infty} \varphi^{\frac{m-2}{2}} e^{-\frac{m\, W\, \varphi}{2(n-m)}}\, d\varphi = 2\left(\frac{n-m}{m\, W}\right)^{\frac{m}{2}} \int_0^{\infty} x^{m-1} e^{-\frac{x^2}{2}}\, d\, x =$$

$$= q_1(m) \left(\frac{n-m}{m\, W}\right)^{\frac{m}{2}}; \tag{8}$$

$q_1(m)$ зависит лишь от числа оцениваемых параметров m и не меняется в процессе отбраковки; символы $q_i(m)$ и в дальнейшем будем использовать для коэффициентов, обладающих этим же свойством. Возвращаясь к (7), получаем для нормирующего множителя

$$Q(n, \mu) = q_2(m) \frac{\Gamma\left(\frac{n-m}{2}\right)}{\int_0^{\mu} t^{\frac{n-m-2}{2}} e^{-t}\, d\, t}. \tag{9}$$

Функция распределения φ имеет вид

$$P(\varphi \leqslant \varphi_{\alpha}) = \frac{2q_2(m)}{(n-m)^{m/2}} \times$$

$$\times \frac{\int_0^{\varphi_{\alpha}} \varphi^{\frac{m}{2}-1} \left(1+\frac{\varphi_m}{(n-m)}\right)^{-\frac{n}{2}} \int_0^{\left(1+\frac{\varphi_m}{n-m}\right)\mu} t^{\frac{n}{2}-1} e^{-t}\, d\, t\, d\varphi}{\int_0^{\mu} t^{\frac{n-m-2}{2}} e^{-t}\, d\, t}. \tag{10}$$

Плотность вероятности определим как

$$\lim_{\varphi_\alpha \to 0} \frac{d P (\varphi \leqslant \varphi_\alpha)}{d V},$$

где V — объем доверительного эллипсоида;

$$V = q_3 (m) (\det K_{\underset{A}{\Delta}} S^{2m} \varphi_\alpha^m)^{1/2}. \tag{11}$$

Очевидно, что

$$\frac{d P}{d V} = \frac{d P}{d \varphi_\alpha} \frac{d \varphi_\alpha}{d V} \tag{12}$$

Из (10) и (11) вытекает:

$$\frac{d P}{d \varphi_\alpha} = q_4 (m) \frac{\varphi_\alpha^{\frac{m}{2}-1} \int_0^{\left(1+\frac{m \varphi_\alpha}{n-m}\right) \mu} t^{\frac{n}{2}-1} e^{-t} d t}{(n-m)^{m/2} \left(1+\frac{m}{n-m} \varphi_\alpha\right)^{n/2} \int_0^{\mu} t^{\frac{n-m}{2}-1} e^{-t} d t};$$

$$\frac{d \varphi_\alpha}{d V} = q_5 (m) \frac{\varphi_\alpha^{\frac{2-m}{2}}}{S^m (\det K_{\underset{A}{\Delta}})^{1/2}}.$$

Используя эти соотношения в (12), приходим к выражению для критерия

$$\lim_{\varphi_\alpha \to 0} \frac{d P}{d V} = q_6 (m) \frac{\int_0^{\mu} t^{\frac{n}{2}-1} e^{-t} d t}{(n-m)^{\frac{m}{2}} S^m (\det K_{\underset{A}{\Delta}})^{1/2} \int_0^{\mu} t^{\frac{n-m}{2}} e^{-t} d t}.$$

Отбраковка прекращается в момент достижения этим функционалом максимума.

Перейдем теперь к более сложному случаю, когда дисперсия заранее неизвестна. Построим вначале алгоритм, рассчитанный на наличие не более одного выброса. В качестве критерия возьмем статистику $\frac{\delta_{i\max}}{\sqrt{S_1^2}}$. Здесь

$$S_1^2 = \frac{\Delta C^T K_1^{-1} \Delta C_1}{n-m-1},$$

т. е. оценка дисперсии, получаемая при изъятии из выборки i-го измерения. Имеется в виду, что и сама точечная оценка вектора $\hat{A}$ построена без участия этого измерения. В данных условиях отношение $\frac{\delta_i}{\sqrt{S_1^2}}$ является дробью Стьюдента и вероятность α_0 того, что

$$\frac{\delta_{i\max}}{S_1} \geqslant t_{n-m-1}, \tag{13}$$

определяется из условия $\alpha_0 = 1 - (1-\alpha)^{n-m-1}$; а t_{n-m-1} пред-

ν	α_0			ν	α_0		
	20 %	15 %	10 %		20 %	15 %	10 %
5	2,6840	2,948	3,325	24	2,830	2,963	3 146
6	2.6805	2,9175	3,255	25	2.838	2.969	3,149
7	2,6837	2,902	3.210	26	2,845	2,9765	3 154
8	2,6897	2.893	3,175	27	2,852	2,9826	3,158
9	2.6980	2,889	3,155	28	2,859	2,988	3,161
10	2,7070	2,8895	3,140	29	2,865	2 9935	3,165
11	2,7160	2,8913	3 133	30	2,872	2,9990	3 169
12	2,724	2 895	3,127	31	2,878	3,0045	3,174
13	2.733	2.8995	3,124	32	2,885	3,0095	3,177
14	2,743	2.9045	3,1216	33	2,890	3,015	3,181
15	2,754	2,9097	3,122	34	2,897	3,020	3,1855
16	2,764	2 9153	3,123	35	2,903	3,025	3,189
17	2 773	2,9218	3 124	36	2,909	3,031	3,193
18	2,782	2,9277	3,126	37	2,915	3,0355	3,197
19	2,791	2,9340	3,129	38	2,920	3,040	3,201
20	2,799	2,9405	3,132	39	2,931	3,050	3,208
21	2,808	2,946	3,135	40	2,979	3,090	3,245
22	2 815	2,953	3,139	75	3,070	3,180	3,319
23	2,823	2,958	3,142	100	3,138	3,240	3,390

ставляет $\alpha/2$-квантиль Стьюдента. В таблице приведены значения для $\alpha/2$-квантилей Стьюдента, соответствующие трем фиксированным уровням значимости α_0, равным 10, 15 и 20% для различного числа степеней свободы. Измерение исключается из обработки при соблюдении (11). Критерий является равномерным, наиболее мощным, несмещенным [6].

Этот алгоритм можно использовать также и при наличии нескольких выбросов, если их величины подчиняются определенным соотношениям. Для иллюстрации такого подхода рассмотрим два выброса. Допустим, что первый выброс отбракован верно. В этом случае приходим к описанной выше модели и отбраковка второго выброса осуществляется согласно (13) при $n-m-2$ степенях свободы. Уточним теперь условия отбраковки первого выброса. Поскольку мы работаем с соотношением (13), его величина должна подчиняться требованию

$$|\delta_{\max 1}| \geqslant t_{n-m-1} S_1$$

или

$$|\delta_{\max 1}| \geqslant t_{n-m-1} \sqrt{\frac{S_2^2 (n-m-2) + \delta^2_{\max 2}}{n-m-1}}. \qquad (14)$$

Алгоритм можно использовать при большем числе выбросов, если каждый последующий по величине будет удовлетворять соотношению, аналогичному (14) для значений S_k^2 и $\delta^2_{\max k}$, соответствующих k-му шагу. Опасным для данного способа отбраковки является близость по модулю нескольких δ_i. Рис. 1—2 иллюстрируют соотношения между величинами возможных выбросов (δ), выявляемых с помощью предлагаемого алгоритма. По оси абсцисс откладывается число степеней свободы $n-m-1$; следовательно, число

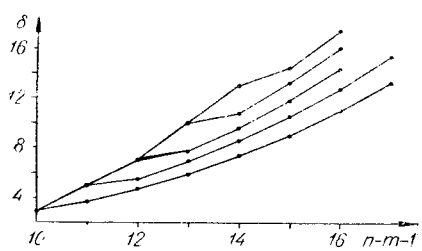

Рис. 1.

достоверных измерений (не пораженных выбросами) принимается равным соответственно $11+m$ (см. рис. 1), $21+m$ (см. рис. 2). Графики построены для 15% уровня значимости α_0. Единицей измерения выбросов принята S, построенная на базе достоверных измерений; практиче-

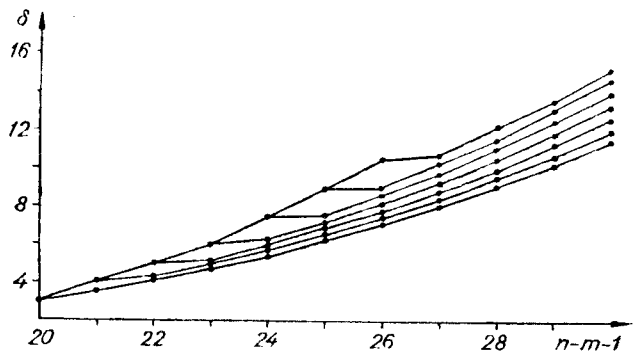

Рис. 2.

ски можно считать, что их модули заданы в σ. Алгоритм работает правильно в том случае, когда последующий (т. е. больший по модулю) выброс лежит не ниже кривой, проходящей через предшествующий выброс. Ветвления графиков соответствуют ситуации, когда выбросы оказываются больше предусмотренных соответствующей кривой. Это влечет естественное увеличение модуля последующих выбросов. Работа алгоритмов проверялась на моделях.

ЛИТЕРАТУРА

1. Н. Г. Микешина. Выявление и исключение аномальных значений (обзор). — Заводская лаборатория, 1966, 32, № 3.
2. G. C. Tiao and Irwin Guttman. Analysis of Outliers with Adjusted Residuals. — Technometrics, 1967, v. 9, № 4.
3. T. I. Anscombe. Rejection of outliers.— Technometrics, 1960, 11, v. 2, № 2.
4. G. E. P. Box and G. C. Tiao. A Bayesian Approach to Some Outlier Problems. — Biometrica, 1968, 55, 1.
5. В. П. Зеленецький. Применение методов теории статистических решений при исключении аномальных измерений.— Изв. АН СССР, Техническая кибернетика, 1969, № 2.
6. Э. Леман. Проверка статистических гипотез. М., «Наука», 1964.
7. Д. К. Фадеев, В. Н. Фадеева. Вычислительные методы линейной алгебры М.—Л., Физматгиз, 1963.
8. С. Уилкс. Математическая статистика. М., «Наука», 1967.

Поступила в редакцию
8 июля 1970 г.