УДК 519.233.22

П. О. Смирнов, Г. Л. Шевляков

ПРИБЛИЖЕНИЕ ОЦЕНКИ Q_N ПАРАМЕТРА МАСШТАБА С ПОМОЩЬЮ БЫСТРЫХ *M*-ОЦЕНОК

Рассмотрены популярные робастные оценки параметра масштаба, обладающие высокой эффективностью, в частности оценка Q_n. Предложено параметрическое семейство М-оценок, обеспечивающих более быстрые вычисления с допустимым снижением пороговой точки. Приведены результаты статистического моделирования.

Ключевые слова: робастность, М-оценки, параметр масштаба.

Один из подходов к оптимизации и принятию решений в условиях неопределенности вероятностных моделей сигналов и помех связан с использованием робастных статистических методов, обеспечивающих устойчивость и надежность результатов статистического анализа к возможным отклонениям от принятых гипотез о распределениях [1]. В свою очередь задачи робастного оценивания параметра масштаба распределений занимают второе по значению место после задач робастного оценивания параметра положения распределений [2]. В данной статье предлагаются быстрые высокоэффективные и робастные оценки параметра масштаба симметричных распределений.

Оценкой масштаба называется любая положительная статистика S_n, которая удовлетворяет равенству $S_n(ax_1, \ldots, ax_n) = aS_n(x_1, \ldots, x_n)$ для $a > 0$ [2–4]. Такая оценка определяет степень разброса значений в выборке и может использоваться как в непараметрических задачах, так и для оценивания значений неизвестного параметра масштаба семейства распределений вероятности. Примером такого параметра может служить параметр σ нормального распределения с плотностью вероятности $N(0, \sigma^2) = \exp(-x^2/2\sigma) / (2\pi\sigma^2)^{1/2}$, а возможной оценкой – стандартное отклонение *SD*.

Эта часто используемая оценка обладает серьезным недостатком: при небольших отклонениях от предполагаемой модели результат может оказаться далеким от истинного. Такими отклонениями могут быть загрязненные данные или ошибочные предположения о законе распределения генеральной совокупности. Классическая статистика рассматривает идеализированные условия, но, согласно [3], в статистических данных, как правило, встречается от 1 до 10 % больших ошибок, а некоторые ряды измерения после тщательной проверки не подтверждают принадлежность нормальному распределению, имея более тяжелые «хвосты».

Подобные отклонения от идеальной модели рассматриваются в робастной статистике. В зависимости от задачи в качестве робастных альтернатив для оценивания масштаба чаще всего используют межквартильный размах $IQR = F^{-1}(3/4) - F^{-1}(1/4)$ или абсолютное медианное отклонение от медианы $MAD = \mathrm{med}_i \mid x_i - \mathrm{med}\, x \mid$. Одним из критериев робастности в данном случае служит пороговая точка оценки, т. е. наибольший процент наблюдений, который можно заменить произвольно большими значениями, оказав лишь ограниченное влияние на значение самой оценки [2]. По этому критерию абсолютное медианное отклонение является более предпочтительным, поскольку оно имеет максимально возможную пороговую точку 50 %. Пороговые точки стандартного отклонения и межквартильного размаха равны 0 и 25 % соответственно.

Стандартное отклонение *SD* имеет минимальную возможную дисперсию для нормального распределения. Робастные же оценки обладают более высокой пороговой точкой, как правило за счет большей дисперсии, т. е. меньшей точности. В этом случае эффективность оценки *MAD* равна всего 36,7 %, поэтому возникает необходимость построения более эффективных оценок с максимальной пороговой точкой.

В работе [5] были предложены такие оценки, одна из которых, а именно Q_n, впоследствии стала часто использоваться на практике. Оценка Q определяется как первый квартиль расстояний между наблюдениями: $Q = \{| x_i - x_j |\}_{(k)}$, $k = C(h, 2)$, $h = [n/2] + 1$, и имеет асимптотическую эффективность 82,3 % и пороговую точку 50 %. Серьезным недостатком является большая вычислительная сложность такой оценки, так как задействовано n^2 разностей между парами значений. Отметим, что даже более эффективный алгоритм [6] требует в 3–5 раз больше времени, чем *MAD*.

В данной статье используется параметрическое семейство оценок масштаба, имеющих такую же или большую эффективность за счет снижения пороговой точки, которая, тем не менее, остается в разумных пределах. Введенный параметр позволяет сохранить баланс между пороговой точкой и эффективностью в зависимости от решаемой задачи.

Постановка задачи. В теории робастности важным инструментом для анализа оценок является подход на основе функции влияния. Функция влияния $IF(x; S, F)$ оценки S на модельном распределении F показывает устойчивость оценки к большим ошибкам в точке x и определяется как ее производная по Гато [3]. Построим оценку масштаба с функцией влияния, совпадающей с $IF(x; Q, F)$, тем самым обеспечив совпадение выражающихся через нее характеристик, в частности асимптотической дисперсии и эффективности.

Рассмотрим класс *M*-оценок масштаба *S*, задаваемых неявным уравнением

$$\sum \chi(x_i / S) = 0,$$

где χ – некоторая оценочная функция, обычно четная и неубывающая при $x > 0$. Выбирая вид функции χ, можно получать различные как робастные, так и неробастные оценки масштаба. Этот класс был введен Хьюбером и подробно рассмотрен в [2; 3].

Известно, что функция влияния таких оценок с точностью до нормирующего множителя совпадает с выбранной оценочной функцией $IF(x; S, F) \propto \chi(x)$. Воспользуемся этим фактом и построим M-оценку масштаба M_α, приближающую Q, положив

$$\chi_\alpha(x) = c_\alpha - \alpha^{-1}(\Phi(x+\alpha) - \Phi(x-\alpha)), \quad \alpha > 0,$$

где $\Phi(x)$ – стандартное нормальное распределение; c_α выбирается из условия состоятельности оценки $\int \chi(x)\, d(x) = 0$. При $\alpha = \Phi^{-1}(5/8)\ \sqrt{2}\ = 0{,}450\ 6$ функция влияния соответствующей оценки M_α будет совпадать с $IF(x; Q, \Phi)$ [5].

Основной результат. Для удобства будем рассматривать другое параметрическое семейство, взяв первые несколько членов разложения $\Phi(x)$ в ряд Тейлора:

$$\chi_\alpha(x) = c_\alpha - \frac{1}{3}\ (6 - \alpha^2\,(x^2 - 1))\ \varphi(x),\ c_\alpha = \\ = (12 - \alpha^2) / (12\sqrt{\pi}).$$

Такое представление позволит получить результат в явном виде через элементарные функции.

Из-за своей простоты также будет интересен частный случай при $\alpha = 0$:

$$\chi_0(x) = 1 / \sqrt{\pi} - 2\varphi(x).$$

Исследования показали, что не имеет смысла рассматривать разложение с большей точностью: выигрыш в характеристиках минимален, а объем вычислений возрастает. В качестве альтернативы можно взять функцию $\chi_{\alpha,\beta}(x)$ с произвольным полиномом второй степени от x^2, но в свете полученных результатов это также представляется избыточным.

Получим характеристики предложенной оценки. Формула для асимптотической дисперсии M-оценок масштаба выглядит следующим образом [3]:

$$V(M_\alpha, \Phi) = \int IF^2(x; M_\alpha, \Phi) d\Phi(x) = \frac{\int \chi_\alpha^2(x)\varphi(x)dx}{(\int x\chi'_\alpha(x)\varphi(x)dx)^2},$$

что дает выражение для эффективности

$$e_\alpha = \frac{1}{2V(M_\alpha, \Phi)} = \\ = \frac{81(\alpha^2-4)^2}{8(432(2\sqrt{3}-3) - 24(8\sqrt{3}-9)\alpha^2 + (16\sqrt{3}-9)\alpha^4}.$$

Максимальная достижимая эффективность составляет 95,9 %, но даже при $\alpha = 0$ она не опускается ниже уровня 80,8 % (рис. 1).

Пороговая точка оценки определяется соотношением

$$\varepsilon^*_\alpha = \frac{-\chi_\alpha(0)}{\chi_\alpha(\infty) - \chi_\alpha(0)} = \frac{12(\sqrt{2}-2) - (\sqrt{2}-4)\alpha^2}{4(\alpha^2-6)}$$

в случае четной монотонно возрастающей при $x > 0$ функции $\chi(x)$. Эти условия будут выполнены при $[0;\ \sqrt{2}]$, причем при нуле достигается максимум $\varepsilon^*_\alpha = 29{,}3$ % (рис. 2).

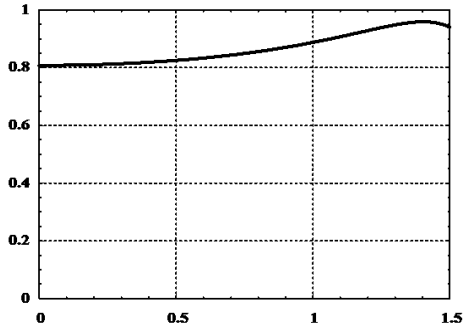

Рис. 1. Эффективность оценки в зависимости от параметра α

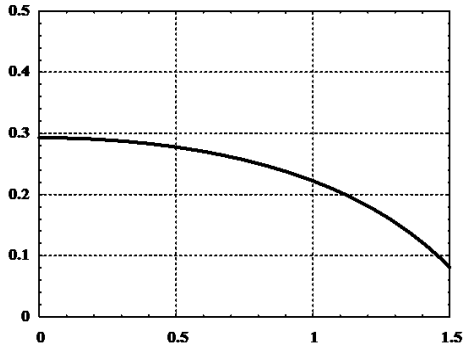

Рис. 2. Пороговая точка оценки в зависимости от параметра α

Вычисление оценки как решения неявного уравнения в большинстве случаев затруднительно, но при этом возможно применение итеративных схем. В частности, можно ограничиться первой итерацией, получив так называемую M-оценку [2]:

$$S_n^{(1)} = S_n^{(0)} - \frac{\sum \chi(x_i / S)}{\left.\frac{\partial}{\partial S}\sum \chi(x_i / S)\right|_{S=S_n^{(0)}}}.$$

Начальное приближение должно быть само по себе в высшей степени робастно. Подставляя предложенную оценочную функцию, получаем, что одношаговая оценка при $\alpha = 0$ задается как уточнение медианы абсолютных отклонений:

$$M_0 = 1{,}483 MAD \cdot \left(1 - \frac{Z_0 - n/\sqrt{2}}{Z_2}\right),$$

$$Z_k \sum u_i^k e^{-u_i^2/2}, \quad u_i = \frac{x_i - \operatorname{med} x}{1{,}483 MAD}.$$

Множитель перед MAD необходим для того, чтобы обеспечить состоятельность оценки на нормальном распределении.

Следует отметить, что функция влияния, а значит асимптотическая дисперсия и эффективность одношаговой оценки, будут несколько отличаться от первоначальных теоретических результатов. Но важным свойством таких оценок является то, что они наследуют пороговую точку начального приближения [6]. При выборе в качестве основы медианы абсолютных отклонений пороговая точка одношаговой оценки M_α повышается до 50 %.

Сравнение оценок. Приведем численные результаты статистического моделирования оценок при 50 000 испытаний на стандартном нормальном распределении (табл. 1). Параметр α имеет значение 0,450 6, что обеспечивает совпадение основных асимптотических характеристик оценок M_α и Q. Время вычисления оценок в миллисекундах соответствует конфигурации Intel Core i7 с частотой 2,8 ГГц. Влияние отклонений от идеальных условий проверялось на модели больших ошибок Тьюки вида $(1 - \varepsilon)\ \Phi(x) + \varepsilon\ \Phi(x/\sigma)$ при $\varepsilon = 0{,}1$, $\sigma = 3$. Результаты моделирования представлены в табл. 2. Таким образом, предложена оценка масштаба, имеющая такую же функцию влияния, асимптотическую дисперсию и эффективность, что и часто используемая оценка Q_n. Результаты моделирования показывают, что одношаговый алгоритм вычисления оценки не только требует значительно меньшего времени, но и обеспечивает меньшее смещение относительно оцениваемой величины, особенно на малых выборках. При подстановке медианы абсолютных отклонений в качестве начального приближения пороговая точка имеет максимально возможное значение 50 %, а свободный параметр позволяет повысить эффективность оценки до 95 %.

Библиографические ссылки

1. Цыпкин Я. З. Информационная теория идентификации. М. : Наука, 1995.

2. Хьюбер Дж. П. Робастность в статистике : пер. с англ. М. : Мир, 1984.

3. Робастность в статистике. Подход на основе функций влияния : пер. с англ. / Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль. М. : Мир, 1989.

4. Шуленин В. П. Введение в робастную статистику. Томск : Изд-во Том. ун-та, 1993.

5. Rousseeuw P. J., Croux C. Alternatives to the median absolute deviation // J. of the American Statistical Association. 1993. Vol. 88, № 424. P. 1273–1283.

6. Rousseeuw P. J., Croux C. The bias of k-step M-estimators // Statistics & Probability Letters. 1994. Vol. 20, № 5. P. 411–420.

Таблица 1

Математическое ожидание и стандартизованная дисперсия оценок на нормальном распределении в зависимости от размера выборки

n	Среднее				Дисперсия				Время, мс			
	SD	*MAD*	*Q*	M_α	*SD*	*MAD*	*Q*	M_α	*SD*	*MAD*	*Q*	M_α
20	0,986	0,958	1,190	0,951	0,532	1,365	0,789	0,656	0,003	0,004	0,007	0,005
60	0,996	0,986	1,062	0,985	0,511	1,350	0,676	0,632	0,010	0,011	0,036	0,014
200	0,998	0,996	1,019	0,995	0,499	1,345	0,638	0,616	0,034	0,034	0,164	0,045
1 000	1,000	0,999	1,004	0,999	0,493	1,364	0,605	0,609	0,168	0,169	1,022	0,228
∞	1,000	1,000	1,000	1,000	0,500	1,361	0,608	0,608	–	–	–	–

Таблица 2

Математическое ожидание и стандартизованная дисперсия оценок на нормальном распределении с загрязнением ($\varepsilon = 10$ %, $\sigma = 3$)

n	Среднее				Дисперсия			
	SD	*MAD*	*Q*	M_α	*SD*	*MAD*	*Q*	M_α
20	1,290	1,047	1,345	1,083	1,601	1,447	0,993	0,886
60	1,324	1,070	1,195	1,110	1,750	1,410	0,875	0,829
200	1,336	1,078	1,144	1,120	1,815	1,426	0,824	0,807
1 000	1,340	1,081	1,126	1,124	1,823	1,415	0,794	0,808

P. O. Smirnov, G. L. Shevlyakov

APPROXIMATION OF THE Q_N-ESTIMATE OF SCALE WITH THE HELP OF FAST *M*-ESTIMATES

Popular highly efficient robust estimates of scale including the Q_n estimate are considered. A parametric family of M-estimates which allows to faster computations with acceptable decrease in breakdown points is offered. The results of the Monte-Carlo simulation are given.

Keywords: robustness, M-estimates, scale parameter.