

В. Л. ЧЕЧУЛИН¹, В.И. ГАЦИЛЁВ¹

¹ Пермский государственный национальный исследовательский университет, Пермь, Российская Федерация

ПРИМЕР ТЕСТИРОВАНИЯ АЛГОРИТМА С ИТЕРАТИВНЫМ ОПРЕДЕЛЕНИЕМ ВЕСОВ

Аннотация. В статье приводятся результаты тестирования работы алгоритма устойчивого оценивания, основанного на неравенстве Чебышева, на большом количестве выборок с одномодальным симметричным распределением, с несимметричным шумом. Отличие рассматриваемого способа оценивания, использующего веса наблюдений, от иных способов взвешивания наблюдений, заключается в том, что функция взвешивания строится по аналогии с неравенством Чебышева, которое не зависит от типа распределения. Это означает, что для используемого способа оценивания не требуется априорного знания типа распределения, метод является свободным от распределений. Основным полученный результат указывает на то, что в итеративном алгоритме для использования первого приближения параметра взвешивающей функции пригодны значения величин как обычного (неустойчивого) стандартного отклонения, так и величина означающая точность измерительного инструмента.

Ключевые слова: устойчивое оценивание, неравенство Чебышева, тестирование, взвешивающая функция.

V.L. CHECHULIN¹, V.I. GRATSILEV¹

¹ Perm State University, Perm, Russian Federation

SAMPLE TESTING OF THE ALGORITHM WITH THE ITERATIVE DETERMINATION OF WEIGHTS

Abstract. The article presents the results of testing the algorithm robust estimation based on the Chebyshev inequality, on a large number of samples with unimodal symmetric distribution with an asymmetric noise. Difference of the considered way of estimation, the using weight of supervision, from different ways of weighing of supervision, is that function of weighing is based by analogy with Chebyshev's inequality which doesn't depend on distribution type. It means that for the used way of estimation it isn't required aprioristic knowledge like distribution, the method is free from distributions. The main result indicates that the iterative algorithm using a first approximation to the parameter of the window function suitable values of both conventional (unstable) standard deviation, and value means the accuracy of the measuring tool.

Keywords: robust estimation, the Chebyshev inequality, testing, weighing function.

1. Предисловие

В процессе получения данных возникают ошибки, шумы, а также всякого рода помехи, которые вносят возмущения в оценки параметров выборки. Требуется такие шумы фильтровать. В настоящее время все более широко используются фильтрующие шумы методы в различных приложениях,— робастных методах управления [17], [18], [19], [20] (а также в экономике, см. [14]) Проблему получения устойчивых (не зависящих от шума) оценок выборки одними из первых решали Хьюбер [8], Хампель [7], которые вычисляли устойчивые оценки параметров выборки с помощью функции влияния. Их подходу следовали и другие авторы [1], [5], см. также [6]. У данного подхода есть существенный недостаток — необходимость априорного знания функции распределения исходной выборки.

Во многих прикладных задачах функция распределения — неизвестна, — единственной априорной информацией о выборке является точность измерительных приборов (процесса измерений), как например, при управлении химико-технологическими процессами [13]. Необходимость разработки методов взвешивания наблюдений, для фильтрации шумов, методов свободных от распределений весьма актуальна.

Поэтому предлагается другой метод вычисления устойчивых оценок параметров выборки, основанный на неравенстве Чебышева [11], не требующий знания функции распределения исходных данных. Факт сходимости метода на одномодальных симметричных распределениях был установлен на небольших примерах в статье [12]. Для подтверждения факта сходимости проводятся вычислительные эксперименты на большом количестве выборок с одномодальным симметричным распределением, с несимметричным шумом, т. е. матожидание шума относительно матожидания исходного сигнала не равно нулю, что отличается от стандартных исходных предположений о характере шума [9]¹.

2. Способ оценивания

У реальных данных истинное распределение не известно, поэтому можно говорить, что каждому единичному измерению сопоставляется некая плотность вероятности, ему соответствующая, если известна точность измерительного инструмента. Затем, когда делается второе измерение, эта плотность вероятности корректируется. Используя такое представление о характере вероятностных закономерностей, для оценки плотности вероятности целесообразно использовать неравенство Чебышева, благодаря которому становится возможным конструировать функцию взвешивания. Эта функция даёт аппроксимацию суммарной плотности вероятности в виде веса наблюдения. Вес — это вероятность (если сумма весов равна единице).

В отличие от параметрических методов оценивания, неравенство Чебышева свободно от вида распределения случайной величины, поэтому далее оно используется для построения оценки, использующей взвешивание наблюдений, которая фильтрует отклоняющиеся наблюдения.

По неравенству Чебышева, для случайной величины $X: \Omega \rightarrow \mathbf{R}$, определенной на вероятностном пространстве (Ω, F, \mathbf{P}) , с конечным математическим ожиданием μ и конечной дисперсией σ имеет место соотношение:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \quad (1)$$

То есть в первом приближении фильтрация сильно отклоняющихся наблюдений сводится к тому, что при известной дисперсии выборки (или ее оценке) можно оценить верхнюю границу вероятности сильно отклоняющихся наблюдений и присвоить им эту оценку в качестве веса (меньшего единицы), наблюдениям же, для которых величина правой части неравенства (1) больше единицы, оставить единичный вес.

В отличие от ядерного сглаживания, предложенного Епанечниковым [3], в котором функция сглаживания задается произвольно и требует предварительной оценки параметра масштаба выборки (для определения «ширины» функции сглаживания), в предлагаемом способе взвешивания использовано неравенство Чебышева, не зависящее от распределения. Более того, при известной точности измерительного инструмента h возможно использование этой информации для построения оценки итеративным способом.

Это же отличает предлагаемый способ и от непараметрической статистики Розенблатта–Парзена (см. [4]), — в предлагаемом способе оценивания вид взвешивающей

¹ Более того, медиана шума, относительно исходного сигнала не равна нулю.

функции определён неравенством Чебышева, а взвешивание производится по аналогии с функцией влияния Хампеля [7].

Используя интерпретацию неравенства Чебышева и подход с использованием функций влияния, вводится некоторая «взвешивающая» функция f_0 , обладающая свойствами [11]:

- 1) Симметричности.
- 2) Ограниченности.
- 3) Убывания на бесконечности до 0.

$$f_0(x_1; x_2) = \begin{cases} f_0(x_1 - x_2) = h_1^2 / (x_1 - x_2)^2, & |x_1 - x_2| > h_1; \\ f_0(x_1 - x_2) = (x_1 - x_2) / h_1, & |x_1 - x_2| \leq h_1, \end{cases} \quad (3)$$

где h_1 – интерпретируется как точность измерительного инструмента. Посредством этой функции f_0 определяются веса наблюдений выборки.

Для получения оценок положения и масштаба используется функция влияния (3). Для каждого наблюдения выборки $X - x_i$ определяется его вес ω_i как сумма влияний $f_0(x_i; x_j)$ на наблюдение x_i наблюдений x_j .

$$\omega_i = \sum_{j=1}^n f_0(x_i; x_j) \quad (4)$$

Затем для выборки строятся обычные оценки среднего с весовыми коэффициентами:

$$Mu(X) = \frac{1}{\sum_{i=1}^n \omega_i} \sum_{i=1}^n x_i \cdot \omega_i. \quad (5)$$

Следует отметить, что выражение для стандартного отклонения дает новую оценку точности измерений h_2 :

$$h_2 = Su(X) = \frac{1}{\sum_{i=1}^n \omega_i} \sum_{i=1}^n (x_i - Mu(X))^2 \cdot \omega_i. \quad (6)$$

При этом по результатам вычислительных экспериментов последовательность h_i сходится к некоторой величине h_0 , являющейся некоторым выражением точности произведенного набора измерений.

Мера масштаба (рассеяния) выборки, приближенно совпадающая с обычной оценкой стандартного отклонения, есть сумма квадратов разностей наблюдений, умноженных на веса обоих наблюдений, деленная на сумму произведений весов [10], [16]:

$$SRu(X) = \frac{1}{2K} \sum_i^n \sum_j^n (x_i - x_j)^2 \cdot \omega_i \omega_j, \quad (7)$$

где $K = \sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j$.

Как уже было сказано, в качестве начального приближения итерационный алгоритм вычисления устойчивых оценок [12] может использовать либо точность измерительного инструмента (h_1), либо стандартное отклонение исходных зашумленных данных (σ).

3. Сравнение оценок при различных начальных приближениях

Рассмотрено два варианта тестирования алгоритма:

1) Данные зашумлены равномерным шумом, начальное приближение – точность измерительного инструмента (h_1).

2) Данные зашумлены равномерным шумом, начальное приближение – стандартное отклонение исходных данных (σ).

В обоих вариантах для тестирования 1000 раз генерировалась выборка размером 1000 наблюдений с нормальным распределением $N(0, 1)$ [2], в которую искусственно внесен шум, доля зашумления 10%. Шум распределен равномерно на интервале $[0; 10]$.

На каждом наборе данных вычислялись различные оценки среднего: математическое ожидание, медиана, устойчивое среднее. Результаты тестирования приведены в таблицах 1, 2.

Таблица 1 – Пример вычисленных оценок, когда 90% исходных данных имеют нормальное распределение ($N(0,1)$), а шум равномерное ($R[0,10]$)

№	Мат. ожидание $M[x]$	Медиана $\mu[x]$	Устойчивое среднее $Mu[x]$ (по h_1)	Устойчивое среднее $Mu[x]$ (по σ)
1	0,45715	0,09285	0,07883	0,07885
2	0,48289	0,08972	0,09020	0,09021
3	0,52032	0,17423	0,13146	0,13146
4	0,45934	0,06145	0,07621	0,07623
5	0,48945	0,12707	0,12007	0,12008
6	0,44709	0,08862	0,05525	0,05527
7	0,50095	0,14550	0,10581	0,10582
8	0,49690	0,17956	0,14317	0,14318
9	0,48191	0,08432	0,06548	0,06550
...
1000	0,48139	0,08500	0,08045	0,08046
Среднее:	0,50016	0,13584	0,11775	0,11777

Между средними значениями полученных оценок вычисляется расстояние:

$$E_i = |a - b|,$$

где, a, b – средние значения различных оценок ($M[x_i], \mu[x_i], Mu[x_i]$),

E_1 – расстояние между матожиданием $M[x]$ и медианой $\mu[x]$,

E_2 – расстояние между медианой $\mu[x]$ и устойчивым средним $Mu[x]$,

E_3 – расстояние между матожиданием $M[x]$ и устойчивым средним $Mu[x]$.

Получены следующие результаты:

Таблица 2 – Результаты тестирования

Начальное приближение	
Точность прибора (h_1)	Стандартное отклонение (σ)
$E_1 = 0,364323$ $E_2 = 0,018087$ $E_3 = 0,382410$	$E_1 = 0,364323$ $E_2 = 0,018087$ $E_3 = 0,382393$

«Треугольники» наглядно демонстрируют, что устойчивые оценки по устойчивости сравнимы с медианой.

Алгоритм с итеративным определением весов протестирован на выборке большого размера с симметричным распределением. Была сгенерирована выборка 1000 раз, на разных выборках оба типа устойчивых оценок дают одинаковые результаты, близкие к медиане. Генерация не зависит от начального задания: h_1 или σ . Результаты совпадают и по критерию Хи-квадрат (значение критерия 1,000000) и по t-тесту Стьюдента (значение критерия 0,990889), что показывает равнозначность выбора начального приближения.

4. Заключение

В данной статье рассмотрено сравнение выбора начального приближения для метода устойчивого оценивания, основанного на неравенстве Чебышева. На множестве одномодальных симметричных выборок с несимметричным шумом сделан вывод об инвариантности рассматриваемого метода относительно выбора начального приближения параметра взвешивающей функции (точность измерительного инструмента или предварительная, неустойчивая, оценка стандартного отклонения), а также о том, что этот метод по устойчивости сравним с медианой.

Допустимость использования в качестве начального приближения масштаба точности измерительного инструмента позволяет модифицировать устойчивые регрессионные оценки [15], используемые для управления технологическими процессами [13].

ЛИТЕРАТУРА

1. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ / пер. с болг. Ю. П. Адлер М.: Финансы и статистика, 1987. 239 с.
2. Генерация стандартного нормального распределения с помощью распределения Бокса-Мюллера. URL: https://ru.wikipedia.org/wiki/Преобразование_Бокса_—_Мюллера

3. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. Т. 14. С. 156–161.
4. Лапко А. В., Лапко В. А. Анализ дисперсия среднеквадратического отклонения аппроксимации непараметрической оценки плотности вероятности ядерного типа // Информатика и системы управления. 2012. №3 (33). С 132-139.
5. Мостеллер Ф., Тьюки Д. Анализ данных и регрессия (в 2-х выпусках) / пер с. англ. М.: Финансы и статистика, 1982. 317+239 с.
6. Справочник по прикладной статистике, в 2-х т. / ред. Ллойд Э., Ледерман У. М.: Финансы и статистика, 1990.
7. Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность в статистике. Подход на основе функций влияния. М.: Мир, 1989. 512 с.
8. Хьюбер Дж. П., Робастность в статистике. М.: Мир, 1984. 304 с.
9. Чеботарев А. С. Способ наименьших квадратов с основами теории вероятностей. М.: Издательство геодезической литературы, 1958. 606 с.
10. Чечулин В. Л. Об оценке масштаба (дисперсии) выборки, не использующей оценку положения (среднего) // Университетские исследования, 2011 (раздел: математика)
URL: http://www.uresearch.psu.ru/files/articles/553_26764.doc
11. Чечулин В. Л. К обоснованию метода устойчивого оценивания посредством неравенства Чебышева // Вестник Пермского университета. Серия: Математика. Механика. Информатика. 2010. Вып. 2 (2). С. 29–32.
12. Чечулин В. Л., Грацилёв В. И. Сходимость оценок масштаба выборки для одномодальных распределений // Университетские исследования, 2014 (раздел: математика). URL: http://www.uresearch.psu.ru/files/articles/713_82346.doc
13. Чечулин В. Л. Метод пространства состояний управления качеством сложных химико-технологических процессов / монография; Перм. гос. нац. исслед. ун-т. Пермь, 2011. 114 с.
14. Чечулин В. Л., Грацилёв В. И. Оценка относительной величины заработной платы с использованием устойчивого оценивая // Университетские исследования, 2013 (раздел: экономика). URL: http://www.uresearch.psu.ru/files/articles/639_58093.doc
15. Чечулин В. Л., Грацилёв В. И. Устойчивое регрессионное оценивание, основанное на неравенстве Чебышева // Университетские исследования, 2013 (раздел: математика)
URL: http://www.uresearch.psu.ru/files/articles/649_93562.doc
16. Чечулин В. Л. О взвешенной оценке масштаба (дисперсии) выборки, не использующей оценку положения (среднего) // Университетские исследования, 2012 (раздел: математика). URL: http://www.uresearch.psu.ru/files/articles/553_26764.doc
17. Dragan Vasile, Morozaan Toader, Stoica Adrian-Mihail. Stochastic Systems Mathematical Methods of Discrete-Time Linear in Robust Control / London: Springer-Verlag 2011.— 346 p.
18. Hassan Bevrani. Robust Power System Frequency Control / New York, USA: Springer. 2009.— 218 p.
19. Siqueira Adriano A. G., Terra Marco H., Bergerman Marcel. Robust Control of Robots / London: Springer-Verlag 2011,— 228 p.
20. Xanthopoulos Petros, Pardalos Panos M., Trafalis Theodore B. Robust Data Mining / London: Springer-Verlag 2010.— 59 p.

REFERENCES

1. Vuchkov I., Boyadzhieva L., Solakov E. *Prikladnoy lineynyy regressionnyy analiz / per. s bolg. Adler Yu. P.* [Applied linear regression analysis / translated from Bolg. Y. P. Adler]. М.: Finansy i statistika, 1987. 239 p. (In Russian)
2. *Generatsiya standartnogo normal'nogo raspredeleniya s pomoshch'yu raspredeleniya Boksa-Myullera* [Generation of a standard normal distribution using the distribution of the Box-Muller]. Available at: https://ru.wikipedia.org/wiki/Preobrazovanie_Boksa_-_Myullera. (In Russian)

3. Epanechnikov V. A. *Neparametricheskaya otsenka mnogomernoy plotnosti veroyatnosti* [Neparametricheskaya multidimensional assessment of the probability density]. *Teoriya veroyatnosti i ee primeneniya* [Probability theory and its application], 1969. T. 14, pp. 156–161 (in Russian).
4. Lapko A. V., Lapko V. A. *Analiz dispersiya srednekvadraticeskogo otkloneniya approksimatsii neparametricheskoy otsenki plotnosti veroyatnosti yadernogo tipa* [Analysis of variance approximation of the standard deviation of nonparametric estimation of the probability density of the nuclear type]. *Informatika i sistemy upravleniya*, 2012, no. 3 (33), pp. 132-139 (in Russian).
5. Mosteller F., T'yuki D. *Analiz dannykh i regressiya (v 2-kh vypuskakh) / per s. angl.* [Data analysis and regression (2 releases) / J. S. eng.] M.: Finansy i statistika, 1982. 317+239 p. (in Russian).
6. Lloyd E., Lederman U. *Spravochnik po prikladnoy statistike, v 2-kh t* [Handbook of Applied Statistics, 2 Vols]. Moscow, Finance and Statistics Publ., 1990. (In Russian)
7. Khampel' F., Ronchetti E., Rausseau P., Shtael' V. *Robastnost' v statistike. Podkhod na osnove funktsiy vliyaniya* [Robustness in statistics. An approach based on the influence functions]. Moscow, Mir Publ., 1989. 512 p. (In Russian)
8. Kh'yuber Dzh. P., *Robastnost' v statistike* [Robust statistics]. Moscow, Mir Publ., 1984. 304 p. (In Russian)
9. Chebotarev A. S. *Sposob naimen'shikh kvadratov s osnovami teorii veroyatnostey* [Method of least squares with the basics of probability theory.]. Moscow, Publisher geodetic literature, 1958. 606 p. (In Russian)
10. Chechulin V. L. *Ob otsenke masshtaba (dispersii) vyborki, ne ispol'zuyushchey otsenku polozheniya (srednego)* [On assessment scale (dispersion) of the sample that does not use the position estimate (average)]. *Universitetskie issledovaniya* [University study]. 2011 (Section: Mathematics). Available at: http://www.uresearch.psu.ru/files/articles/553_26764.doc (In Russian)
11. Chechulin V. L. *K obosnovaniyu metoda ustoychivogo otsenivaniya posredstvom neravenstva Chebysheva* [Proof of the method of evaluation of sustainable means of Chebyshev inequality]. *Vestnik Permskogo universiteta. Seriya: Matematika. Mekhanika. Informatika*, 2010, no. 2 (2)., pp. 29–32 (in Russian).
12. Chechulin V. L., Gratsilev V. I. *Skhodimost' otsenok masshtaba vyborki dlya odnomodal'nykh raspredeleniy* [Convergence estimates of the scale of sampling for unimodal distributions]. *Universitetskie issledovaniya* [University study]. 2014 (Section: Mathematics). Available at: http://www.uresearch.psu.ru/files/articles/713_82346.doc (In Russian)
13. Chechulin V. L. *Metod prostranstva sostoyaniy upravleniya kachestvom slozhnykh khimiko-tekhnologicheskikh protsessov* [By the state space of quality management of complex chemical-engineering processes]. Perm, Perm. state. nat. Issled. Univ Publ., 2011. 114 p. (In Russian)
14. Chechulin V. L., Gratsilev V. I. *Otsenka otноситel'noy velichiny zarabotnoy platy s ispol'zovaniem ustoychivogo otsenivaya* [Assessment of the relative value of wages using sustainable estimating]. *Universitetskie issledovaniya* [University study]. 2013 (Section: Economy). Available at: http://www.uresearch.psu.ru/files/articles/639_58093.doc (In Russian)
15. Chechulin V. L., Gratsilev V. I. *Ustoychivoe regressionnoe otsenivanie, osnovannoe na neravenstve Chebysheva* [Sustainable regression estimation based on inequality Chebyshev]. *Universitetskie issledovaniya* [University study]. 2013 (Section: Mathematics). Available at: http://www.uresearch.psu.ru/files/articles/649_93562.doc (In Russian)
16. Chechulin V. L. *O vzyeshennoy otsenke masshtaba (dispersii) vyborki, ne ispol'zuyushchey otsenku polozheniya (srednego)* [On a weighted assessment scale (variance) sample that does not use the position estimate (average)]. *Universitetskie issledovaniya* [University study]. 2012 (section: matematika). Available at: http://www.uresearch.psu.ru/files/articles/553_26764.doc (In Russian)
17. Dragan Vasile, Morozan Toader, Stoica Adrian-Mihail. *Stochastic Systems Mathematical Methods of Discrete-Time Linear in Robust Control* / London: Springer-Verlag 2011.— 346 p.
18. Hassan Bevrani. *Robust Power System Frequency Control* / New York, USA: Springer. 2009.— 218 p.

19. Siqueira Adriano A. G., Terra Marco H., Bergerman Marcel. Robust Control of Robots / London: Springer-Verlag 2011,— 228 p.
20. Xanthopoulos Petros, Pardalos Panos M., Trafalis Theodore B. Robust Data Mining / London: Springer-Verlag 2010.— 59 p.

© Чечулин В. Л., Грацилёв В. И., 2015

ИНФОРМАЦИЯ ОБ АВТОРАХ

Чечулин Виктор Львович – старший преподаватель, Пермский государственный национальный исследовательский университет, Пермь, Российская Федерация, e-mail: chechuulinvl@mail.ru

Грацилёв Вадим Игоревич – магистрант Пермский государственный национальный исследовательский университет, Пермь, Российская Федерация; e-mail: Vadim.Gratsilev@yandex.ru

INFORMATION ABOUT AUTHORS

Chechulin Viktor Lvovich – senior lector, Perm State University, Perm, Russian Federation, e-mail: chechuulinvl@mail.ru

Gratile Vadim Igorevich – graduate student, Perm State University, Perm, Russian Federation, e-mail: Vadim.Gratsilev@yandex.ru