

# Исследование одной стохастической оценки центра симметрии в многомерном пространстве

Филатова А.А., СПбГУ, Санкт-Петербург [r ii.filatova@gmail.com](mailto:r ii.filatova@gmail.com),  
Ермаков М.С., СПбГУ, Санкт-Петербург [erm2512@mail.ru](mailto:erm2512@mail.ru)

## Аннотация

Рассматривается новый робастный стохастический алгоритм для оценки центра симметрии многомерных распределений с выпуклыми центрально-симметричными поверхностями уровня плотности. Исследованы его временная сложность, скорость сходимости и робастность. Также рассмотрена модификация алгоритма, повышающая точность оценки. Представлены сравнительные численные эксперименты с существующими алгоритмами оценки многомерного параметра положения, включая медиану Тьюки, Оджа и геометрическую медиану. Поскольку в рассматриваемом классе распределений эти алгоритмы оценивают центр симметрии, сравнение является объективным.

## Введение

В одномерном случае естественные оценки параметра положения (медиана, математическое ожидание) известны и их робастные версии легко определяются. Однако в многомерном случае при построении робастных версий оценок параметра положения статистики сталкиваются с серьезными трудностями. Известно, что не существует естественного обобщения одномерной медианы для многомерных данных [2, 3]. Это связано с тем, что в многомерном пространстве, в отличие от  $\mathbb{R}^1$ , нет естественного упорядочивания точек, что делает прямое перенесение одномерных концепций невозможным. В результате в литературе предлагаются различные подходы к определению многомерной медианы, такие как геометрическая медиана [5], медиана Оджа [6], концепции на основе глубины данных, в частности, медиана Тьюки [1] и другие [3].

Хотя единого общепринятого обобщения одномерной медианы на многомерный случай не существует, есть определенные свойства, которыми оно должно обладать. Например, для одномерных симметричных распределений медиана совпадает с центром симметрии. Естественно предположить, что то же должно быть верно и в многомерном пространстве. В этом случае мы в качестве обобщения понятия симметричного распределения рассмотрим распределения с выпуклыми центрально-симметричными поверхностями уровня плотности. Поэтому основой предлагаемого алгоритма является сообра-

жение, что для подобных распределений он должен оценивать центр симметрии.

Важным практическим ограничением многих существующих методов является их высокая вычислительная сложность [6, 7, 8], что существенно сужает область их применения. Для преодоления этого ограничения в работе предлагается стохастический подход, потенциально обеспечивающий как робастность, так и снижение вычислительных затрат.

## Новый стохастический алгоритм для оценки многомерного параметра положения

### *Предположения алгоритма*

Алгоритм 1 оценивает параметр положения для распределений, поверхности уровня плотности  $f$  которых являются выпуклыми центрально-симметричными, то есть для любого  $\mathbf{x} \in \mathbb{R}^d$ :

- $f(\mathbf{x} + \mathbf{x}_0) = \text{const}$  – выпуклое множество;
- $f(\mathbf{x}_0 + \mathbf{x}) = f(\mathbf{x}_0 - \mathbf{x})$ , где  $\mathbf{x}_0$  – центр симметрии.

В этом случае параметр положения определяется как значение, совпадающее с центром симметрии распределения  $\mathbf{x}_0 \in \mathbb{R}^d$ .

Из предположений следует, что алгоритм 1 может быть рассмотрен как метод оценки центра симметрии. Величина, которая получается в результате работы алгоритма, далее будет называться **стохастической медианой**.

### *Формулировка алгоритма*

Алгоритм 1 сводит многомерную задачу к последовательности одномерных подзадач. На каждом шаге итерационного процесса выбирается случайное направление прямой, затем наблюдения проецируются на нее, и вычисляется медиана для одномерных точек-проекций. Такой подход позволяет последовательно приближаться к центру симметрии, при этом используя простые вычисления и обеспечивая эффективное решение задачи оценки параметра положения в многомерном случае.

### **Алгоритм 1** оценки многомерной медианы

1. Рассматриваем выборку  $\mathbf{x}_1, \dots, \mathbf{x}_n$  из распределения с функцией плотности  $f$ , которая удовлетворяет предположениям алгоритма.

2. Произвольным образом выбираем точку  $\hat{\mathbf{m}}_1$  — начальное приближение. Задаем погрешность вычисления  $\varepsilon$ .
3. Моделируем случайный вектор  $\mathbf{u}_i$ , равномерно распределенный на единичной сфере. Проводим прямую

$$l_i = \{\hat{\mathbf{m}}_i + \lambda \mathbf{u}_i, \lambda \in \mathbb{R}\}.$$

4. Проецируем наблюдения  $\mathbf{x}_1, \dots, \mathbf{x}_n$  на прямую  $l_i$ , получаем точки-проекции  $y_{i1}, \dots, y_{in}$ .
5. Находим медиану  $\hat{\mathbf{m}}_{i+1}$  точек  $y_{i1}, \dots, y_{in}$ .
6. Алгоритм завершается, когда выполняется условие  $\|\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_{i+1}\| < \varepsilon$ . Иначе увеличиваем счётчик шагов  $i$  на 1 и возвращаемся к шагу 3.

Чтобы уменьшить влияние случайности на финальную оценку, рассмотрим **модификацию** алгоритма 1. Вместо того чтобы использовать результат последней итерации, предлагается после достижения заданной точности  $\varepsilon$  сделать еще  $k$  итераций и в качестве финальной оценки взять среднее по этим  $k$  итерациям. Пусть до достижения заданной точности алгоритм 1 выполнялся  $N$  шагов, тогда

$$\hat{\mathbf{m}}_{\text{avg}_k} = \frac{1}{k} \sum_{i=N+1}^{N+k} \hat{\mathbf{m}}_i.$$

Усреднение последних итераций позволяет снизить разброс оценки и делает её более устойчивой.

## Анализ временной сложности алгоритма

Сложность одного шага алгоритма 1 для выборки объемом  $n$  в пространстве размерности  $d$  выглядит следующим образом:

$$O(d) + O(nd) + O(n) + O(d) + O(d) = O(nd). \quad (1)$$

Это означает, что время выполнения одного шага алгоритма 1 линейно зависит от количества точек в выборке  $n$  и размерности пространства  $d$ . Если алгоритм 1 выполняется  $k$  итераций до сходимости, общая временная сложность алгоритма будет  $O(knd)$ , где  $k$  — число итераций.

### ***Сравнение с некоторыми известными алгоритмами оценки медианы***

В таблице 1 представлена сравнительная информация о временной сложности алгоритмов оценки различных обобщений медианы в пространстве размерности  $d$ .

Медиана	Временная сложность	Источник
Тьюки*	$O((d+k)n^2 + n^2 \log n)$	[4]
Оджа	$O(kdn^d \log n)$	[6]
Геометрическая	$O(knd)$	[5]
Стохастическая	$O(knd)$	(1)

Таблица 1: Временная сложность вычисления многомерной медианы

Как видно из таблицы 1, алгоритмы оценки геометрической и стохастической медианы обладают наилучшей асимптотикой, что делает их более предпочтительными для работы с большими выборками ( $n \gg 1$ ) и в высокоразмерных пространствах ( $d \gg 1$ ).

Также отметим, что в таблице 1 указана временная сложность аппроксимационного алгоритма для оценки медианы Тьюки (ABCDepth) [4], так как точный алгоритм для больших  $n$  и  $d$  является  $NP$ -сложной задачей [10]. Однако, используя технику рандомизированной оптимизации [9], можно добиться сложности  $O(n^{d-1})$  для точного вычисления медианы Тьюки.

Для центрально-симметричных выпуклых распределений все рассмотренные медианы совпадают с центром симметрии. Это следует из их определений:

- **Медиана Тьюки** максимизирует глубину Тьюки — минимальное число точек по одну сторону от любой гиперплоскости, проходящей через нее [1]. В случае центральной симметрии максимальная глубина достигается в центре симметрии, так как любое смещение от него её уменьшает.
- **Медиана Оджа** минимизирует сумму объёмов симплексов по всем комбинациям возможных  $d$  точек выборки [6]. При центральной симметрии и выпуклости эта сумма минимизируется в центре симметрии, так как эти свойства гарантируют, что любое смещение увеличивает сумму объёмов.
- **Геометрическая медиана** минимизирует сумму евклидовых расстояний до всех точек выборки [5]. Для центрально-симметричного распре-

деления минимум достигается в центре симметрии, так как симметрия означает, что любое смещение увеличивает сумму расстояний.

### Сходимость алгоритма для случая равномерного распределения в шаре

**Предложение 1.** Пусть  $d \geq 2$  – размерность пространства,  $T_\varepsilon$  – среднее количество шагов алгоритма до попадания в  $\varepsilon$ -окрестность,  $r_0$  – расстояние между начальным приближением и нулем.

Алгоритм 1 в случае равномерного распределения в шаре сходится, причем

$$\frac{T_\varepsilon}{\ln(r_0/\varepsilon)} \xrightarrow{P} \frac{1}{-E_d} \text{ при } \varepsilon \rightarrow 0,$$

$$\text{где } E_d = \begin{cases} -\sqrt{\pi} \cdot \left[ \ln 2 - \sum_{i=1}^{d-2} \frac{(-1)^{i+1}}{i} \right], & d - \text{четное,} \\ \sqrt{2} \cdot \left[ \ln 2 - \sum_{i=1}^{d-2} \frac{(-1)^{i+1}}{i} \right], & d - \text{нечетное.} \end{cases}$$

#### *Идея доказательства.*

Рассмотрим идеальную ситуацию, когда наблюдений очень много и приближения к центру симметрии не зависят от объема и расположения точек выборки.

Пусть  $r(\hat{\mathbf{m}}_i) = r_i$  — расстояние между  $i$ -м приближением к центру симметрии  $\hat{\mathbf{m}}_i$  и истинным значением центра симметрии.

Тогда с помощью геометрических рассуждений можно показать, что последовательность из  $-\ln r_i$  является процессом восстановления. Сходимость этого процесса может быть доказана с помощью элементарной теоремы восстановления [11]. Также из геометрических соображений вычисляется средняя величина шага этого процесса.

Таким образом, имеет место сходимость по вероятности нормированного числа шагов алгоритма к константе.

Из значения константы  $E_d$  следует, что с увеличением размерности пространства  $d$  алгоритм будет сходиться за большее количество шагов, что ожидаемо.

## Исследование сходимости и робастности

### Сходимость

На рисунке 1 представлены зависимости нормы ошибки от количества итераций для равномерного распределения в шаре и для стандартного многомерного нормального в случае размерностей 2-6.

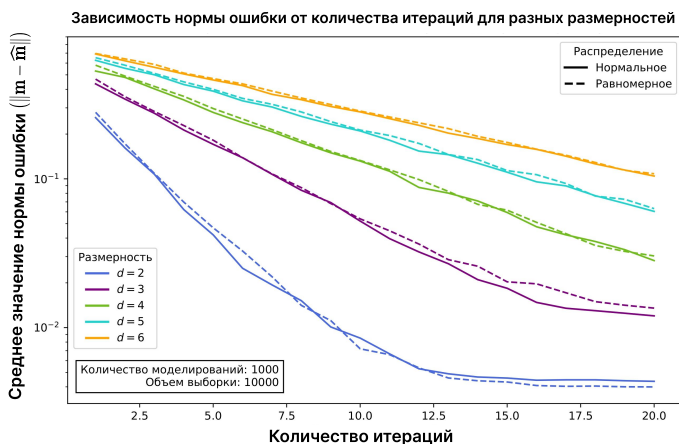


Рис. 1: Зависимость нормы ошибки от количества итераций. Делается 1000 моделирований, чтобы получить среднее значение нормы ошибки.

Результаты моделирования показывают, что при увеличении размерности скорость сходимости алгоритма замедляется и приближение к истинному значению происходит за большее количество итераций. Причем это характерно для обоих рассматриваемых распределений в равной степени. Это может говорить о том, что тип распределения оказывает незначительное влияние на общую тенденцию сходимости алгоритма.

Как видно из рисунка 1, на начальных итерациях скорость сходимости подобна линейной, после чего наблюдается выполаживание. Наиболее отчетливо этот эффект виден для размерностей 2 и 3, тогда как в случае больших размерностей выполаживание происходит за большее число итераций. На основании этого можно предположить, что скорость сходимости остаётся линейной при любой размерности, а выполаживание связано с тем, что выборка имеет конечный объем: на больших итерациях алгоритм может прыгать по элементам выборки, лежащим рядом с центром симметрии.

## Робастность

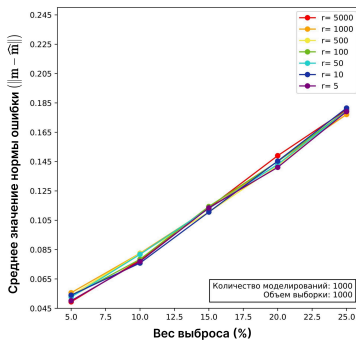
### Случай фиксированной точности

Рассмотрим зависимость нормы ошибки от веса выброса для различных расстояний от истинного значения центра симметрии в случае фиксированной точности для двумерного стандартного нормального распределения. Вес выброса — это доля точек в выборке, которые заменяются на точку-выброс. Результаты моделирования представлены на рисунке 2 (а). Видно, что точность оценки медианы уменьшается практически линейно с увеличением веса выброса. Это означает, что точность пропорциональна весу выброса, и при более значимых выбросах оценка становится менее точной. Причем от удаленности от истинного значения это не зависит, так как для всех  $r$  результаты примерно одинаковы при заданной точности  $\varepsilon$ .

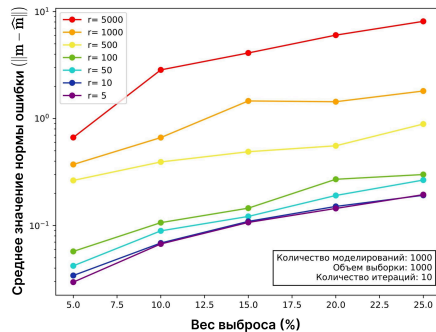
### Случай фиксированного количества итераций

Изменим критерий остановки алгоритма 1: совершим ровно  $k$  итераций вне зависимости от того, достигнута ли точность  $\varepsilon$ , и запустим алгоритм вновь на тех же данных. Результаты моделирования представлены на рисунке 2 (б).

Из результатов эксперимента можно предположить, что удаленность выброса влияет на скорость сходимости алгоритма 1: для меньших значений  $r$  алгоритм сходится быстрее, чем для больших значений  $r$ .



(а) Фиксирована точность



(б) Фиксировано количество итераций

Рис. 2: Зависимость нормы ошибки от веса выброса и удаленности от истинного значения. Часть выборки заменяется на точку с заданным весом на расстоянии  $r \in \{5, 10, 50, 100, 500, 1000, 5000\}$  от истинного значения. Рассматриваются веса 5%, 10%, 15%, 20%, 25% от объема выборки. Алгоритм 1 запускается для  $\varepsilon = 0,01$  (а) и для фиксированного количества итераций  $k = 10$  (б).

## Сравнение результатов работы с другими алгоритмами

Сравним работу предложенного алгоритма с популярными алгоритмами оценивания многомерного параметра положения.

Рассмотрим изменение ошибки оценки центра симметрии для разных объемов выборки в случае многомерного стандартного нормального распределения размерности 4. Результаты моделирования представлены на рисунке 3.

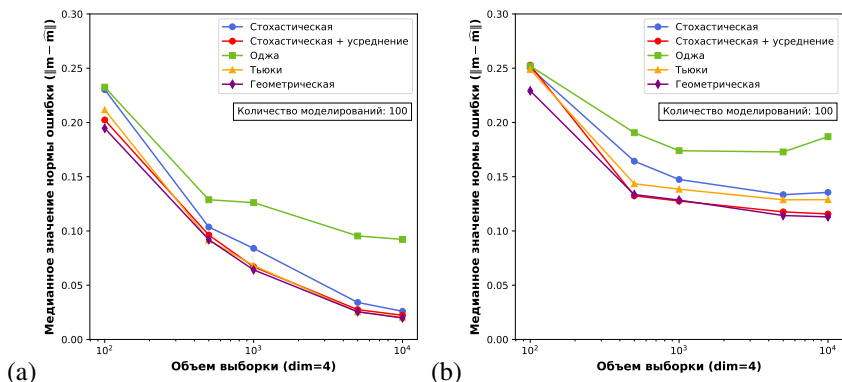


Рис. 3: Медианные значения нормы ошибок для выборок объема 100, 500, 1000, 5000 и 10000 из нормального распределения с (b) выбросами и (a) без них.

Исходя из вида рисунка 3 (a) можно предположить, что алгоритм 1 при заданных объемах выборки сопоставим по точности с геометрической медианой — конкурентом по временной сложности, а также с медианой Тьюки.

Заменим в выборке 5% наблюдений на выброс — точку  $(8, 6, 5, 10)^T$  и запустим алгоритмы вновь. Результаты представлены на рисунке 3 (b). Видно, что предложенный алгоритм 1 уже немного хуже оценивает параметр положения, однако точность все равно близка к геометрической медиане и медиане Тьюки.

Таким образом, результаты сравнения показывают, что новый алгоритм 1 при более подробном исследовании может конкурировать с известными алгоритмами оценки многомерного параметра положения.

## Заключение

В рамках работы был реализован новый стохастический алгоритм робастного оценивания центра симметрии в многомерном пространстве. Исследо-



вана на модельном примере и строго доказана его скорость сходимости, а также временная сложность. Проведено сравнение с распространёнными алгоритмами, которое показало, что предложенный алгоритм и его модификация, повышающая точность оценки, дают сопоставимые результаты и не уступают по точности. Алгоритм эффективен для больших выборок и высоких размерностей, что делает его перспективной альтернативой существующим методам.

### Список литературы

- [1] Rousseeuw P., Struyf A. Computation of robust statistics: depth, median, and related measures // Handbook of Discrete and Computational Geometry. Second ed. – Chapman & Hall/CRC. – 2004. – P. 1279–1292.
- [2] Huber P. J. Robust Statistics // Wiley. – 1981. – P. 308.
- [3] Small C. G. A Survey of Multidimensional Medians // International Statistical Review. – 1990. – Vol. 58, No. 3. — P. 263–277.
- [4] Bogićević M., Merkle M. Approximate Calculation of Tukey's Depth and Median With High-dimensional Data // Yugoslav Journal of Operations Research. – 2018. – Vol. 28, No. 4. – P. 475–499.
- [5] Cohen M. B., Lee Y. T., Miller G., Pachocki J., Sidford A. Geometric median in nearly linear time // Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing. – 2016. – P. 9–21.
- [6] Ronkainen T., Oja H., Orponen P. Computation of the multivariate Oja median // Developments in robust statistics: International Conference on Robust Statistics ICORS '01, Stift Vorau, Itävalta, heinäkuu 2001. – 2002. – P. 344–359.
- [7] Langerman S., Steiger W. Optimization in arrangements // Proc. 20th Sympos. Theor. Aspects Comp. Sci. Berlin: Springer. – 2003. – Vol. 2607. P. 50–61.
- [8] Aloupis G., Langerman S., Soss M., Toussaint G. Algorithms for bivariate medians and a Fermat-Torricelli problem for lines // Comput. Geom. – 2003. – Vol. 26. P. 69–79.
- [9] Chan T.M. An optimal randomized algorithm for maximum Tukey depth // Proc. 15th Annu. ACM-SIAM Sympos. Discrete Algorithms (SODA '04). – 2004. – P. 430–436.

- [10] Dyckerhoff R., Mozharovskyi P. Exact computation of the halfspace depth // Computational Statistics & Data Analysis. – 2016. – Vol. 98. – P. 19–30.
- [11] Боровков А. А. Теория вероятностей. — М.: Эдиториал УРСС. – 1999. – 472 с.