Исследование одной робастной оценки

Филатова Арина Алексеевна, гр. 20.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: д.ф.-м.н., профессор М.С. Ермаков
Рецензент: к.ф.-м.н., с.н.с. В.Н. Солев

Санкт-Петербург, 2024

Введение

Проблема:

- Медленная скорость сходимости детерминированных алгоритмов робастного оценивания многомерного параметра положения.
- Необходимость повышения эффективности.

Актуальность:

- Отсутствие стохастических подходов робастного оценивания многомерного параметра положения.

Цель:

- Исследование нового робастного стохастического алгоритма для оценки многомерного параметра положения.

Постановка задачи

- Реализовать и исследовать новый робастный алгоритм для оценки одного из вариантов многомерного параметра положения.
- Получить некоторые начальные результаты относительно его сходимости, скорости сходимости и робастности.
- Сравнить результат работы предложенного алгоритма с наиболее распространенными алгоритмами для оценки многомерного параметра положения.

Предположения алгоритма

Алгоритм несмещенно оценивает параметр положения для распределений, поверхности уровня плотности f которых являются **выпуклыми центрально-симметричными**, то есть такими, что 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 обычно определяется как значение, совпадающее с центром симметрии $\mathbf{x}_0 \in \mathbb{R}^d$. Поэтому предложенный алгоритм можно рассматривать как оценку некоторого многомерного обобщения медианы.

Алгоритм вычисления исследуемой оценки многомерной медианы [М. Ермаков, 2022]

1. Рассматриваем выборку $\mathbf{x}_1, \ldots, \mathbf{x}_n$ из распределения с функцией плотности f.
2. Произвольным образом выбираем точку $\widehat{\mathbf{m}}_1$ — начальное приближение. Задаем погрешность вычисления ε.
3. Моделируем случайный вектор $\mathbf{u}_i$, равномерно распределенный на единичной сфере. Проводим прямую

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

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

Временная сложность и ее сравнение с некоторыми другими алгоритмами вычисления медианы

- Временная сложность для выборки объемом n в пространстве размерности d для одного шага алгоритма:

$$O(d) + O(nd) + O(n) + O(d) + O(d) = O(nd).$$

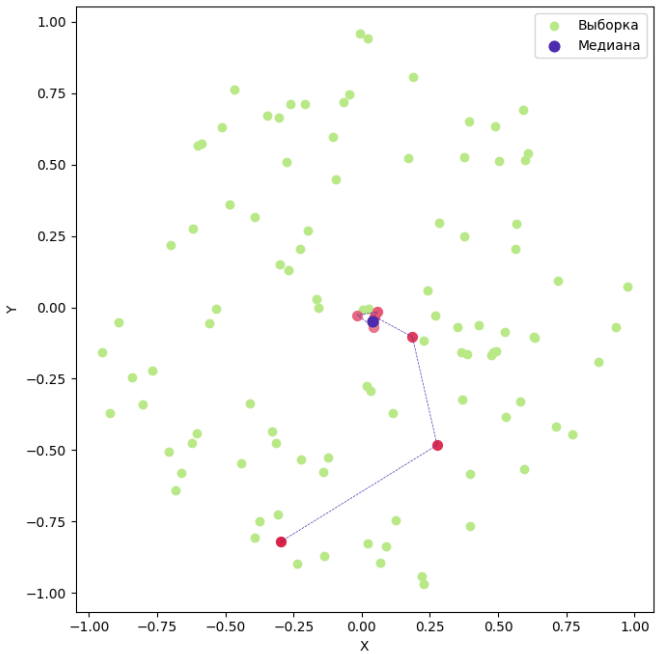
- В таблице представлена сравнительная информация о временной сложности методов вычисления *двумерной* медианы.

Медиана	**Временная сложность**	**Источник**
Тьюки	$O(n \log^3 n)$	[Langerman, 2003]
Геометрическая*	$O(n)$	[Cohen, 2016]
Оджа	$O(n \log^3 n)$	[Aloupis, 2003]
Стохастическая*	$O(n)$	–

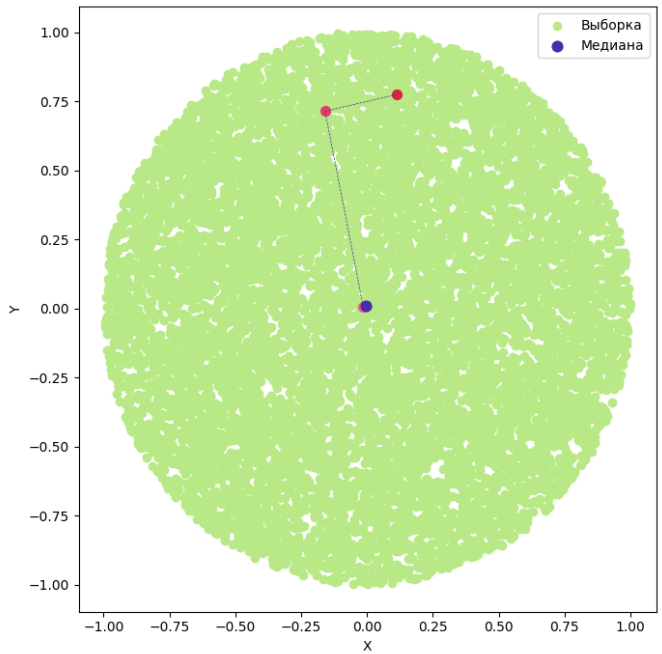
Результат работы алгоритма

Мной была написана функция на языке программирования Python, которая реализует данный алгоритм.

В результате ее работы для точек, равномерно распределенных в единичном круге, получается следующий результат.



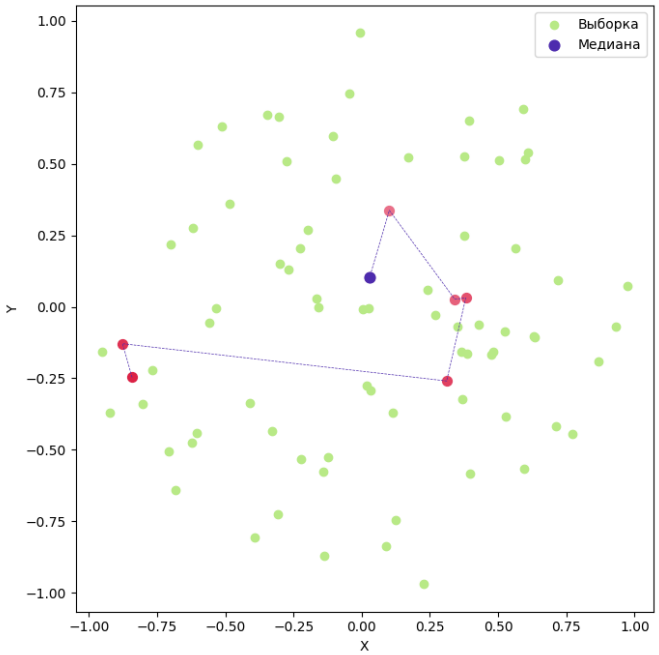
100 точек



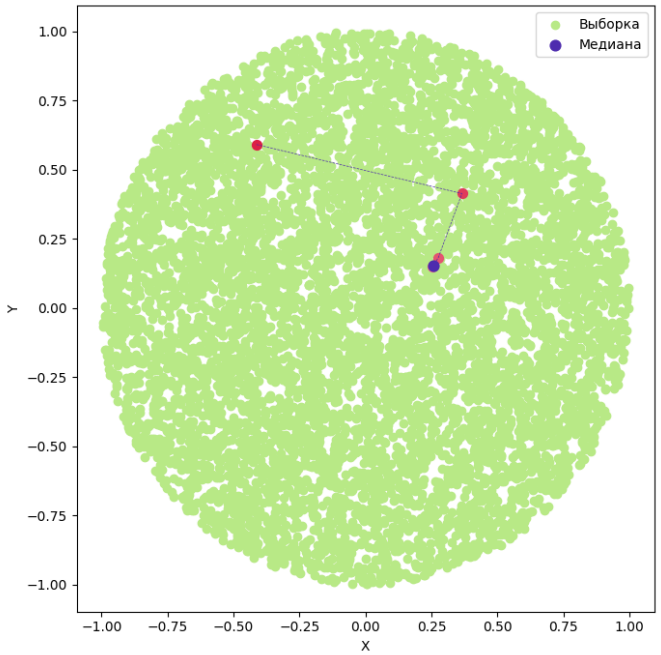
10000 точек

Результат работы алгоритма

При аналогичных условиях, если в выборке заменить 25% точек на точку-выброс $(3,\ 4)^{\mathbf{T}}$, результат следующий.



100 точек



10000 точек

Исследование сходимости и скорости сходимости в случае равномерного распределения в шаре

Мной был получен следующий теоретический результат.

Предложение

Пусть $d \geqslant 2$ – размерность пространства, T_ε – количество шагов алгоритма до попадания в ε-окрестность, r_0 – расстояние между начальным приближением и нулем.

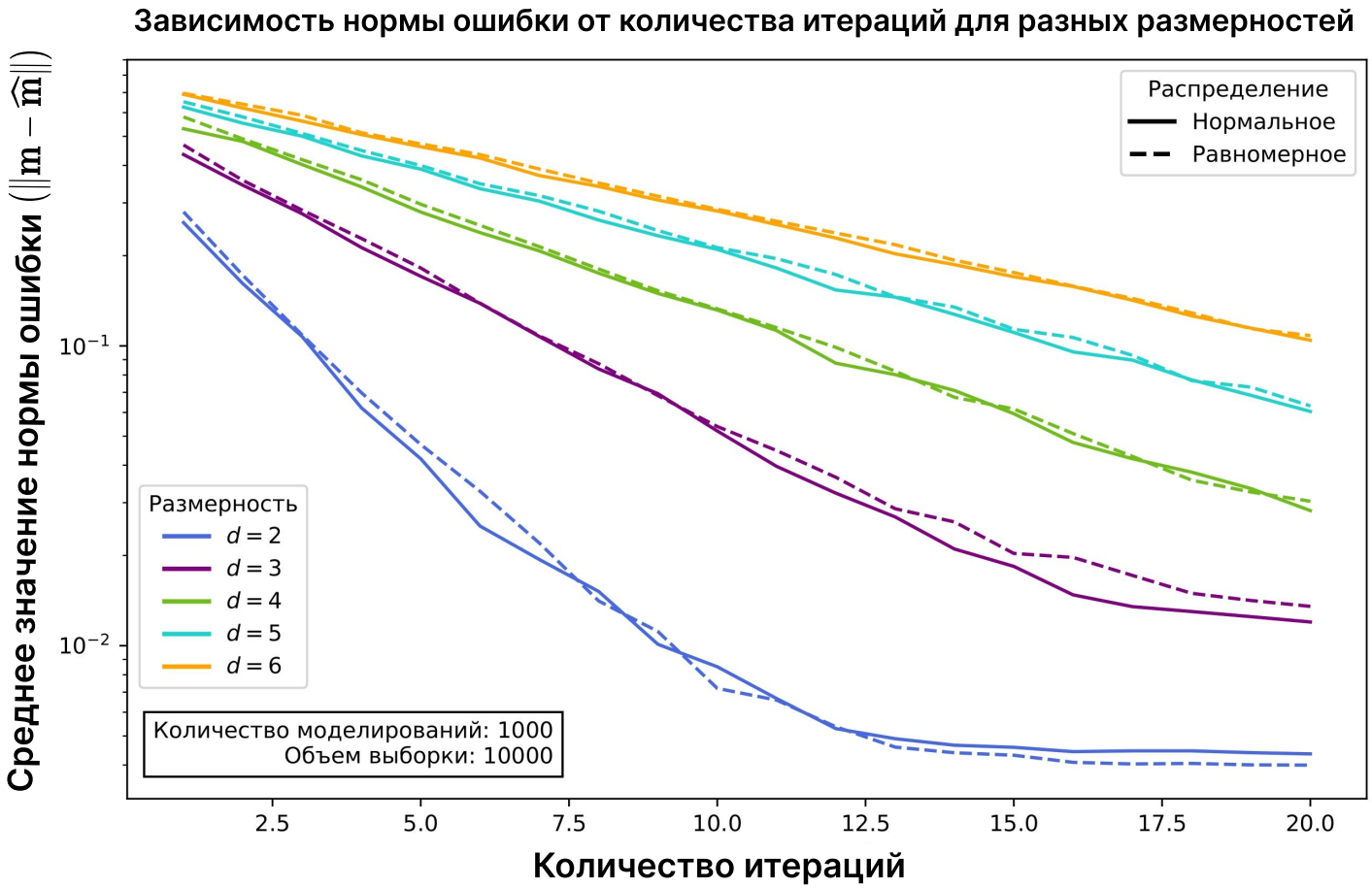
Алгоритм в случае равномерного распределения в шаре сходится, причем

$$\frac{T_\varepsilon}{\ln(r_0/\varepsilon)} \xrightarrow{\mathrm{P}} \frac{1}{-\mathrm{E}_d} \text{ при } \varepsilon \to 0,$$

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

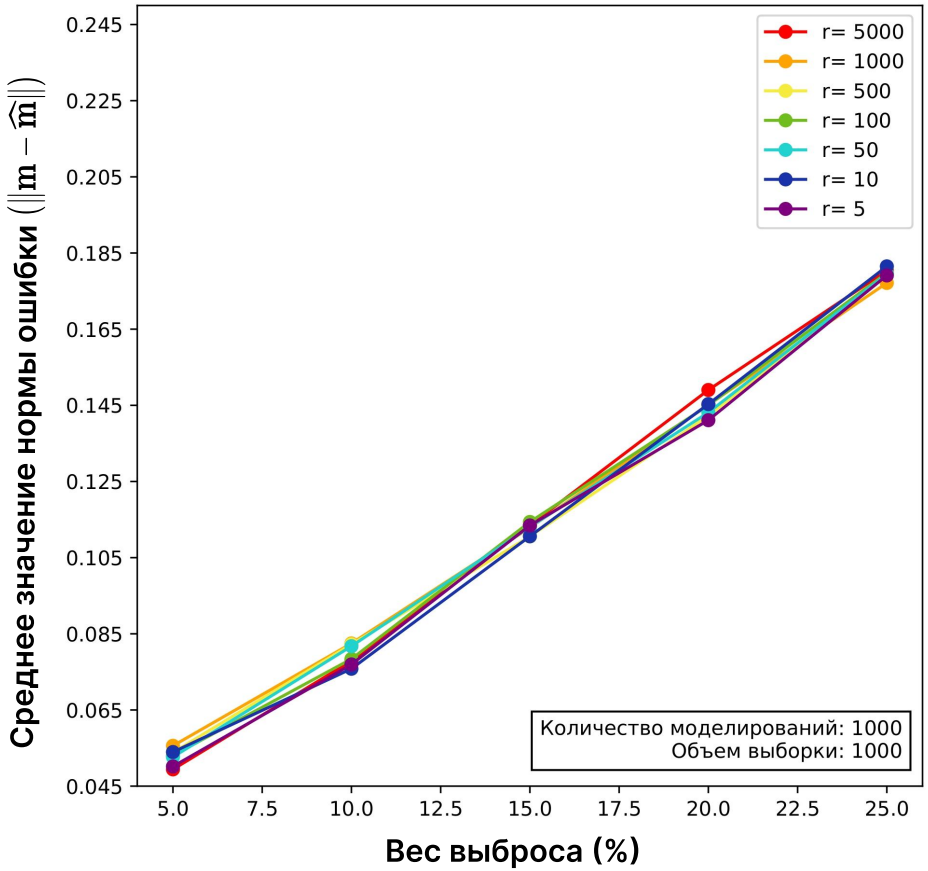
Исследование скорости сходимости в зависимости от размерности

Рассмотрим график зависимости нормы ошибки от количества итераций.



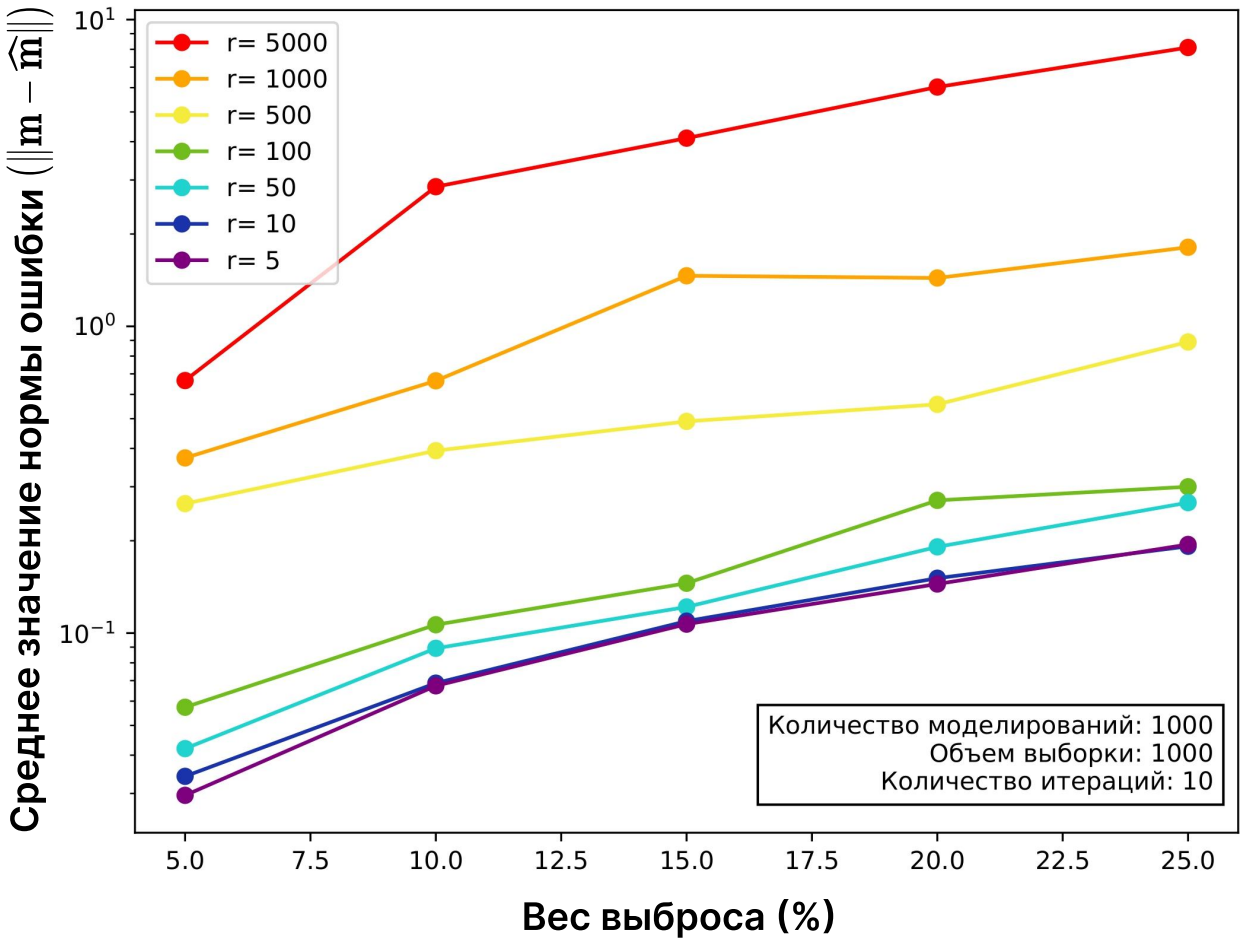
Исследование робастности

Рассмотрим график зависимости нормы ошибки от веса выброса для различных расстояний от истинного значения в случае фиксированной точности для двумерного нормального распределения.



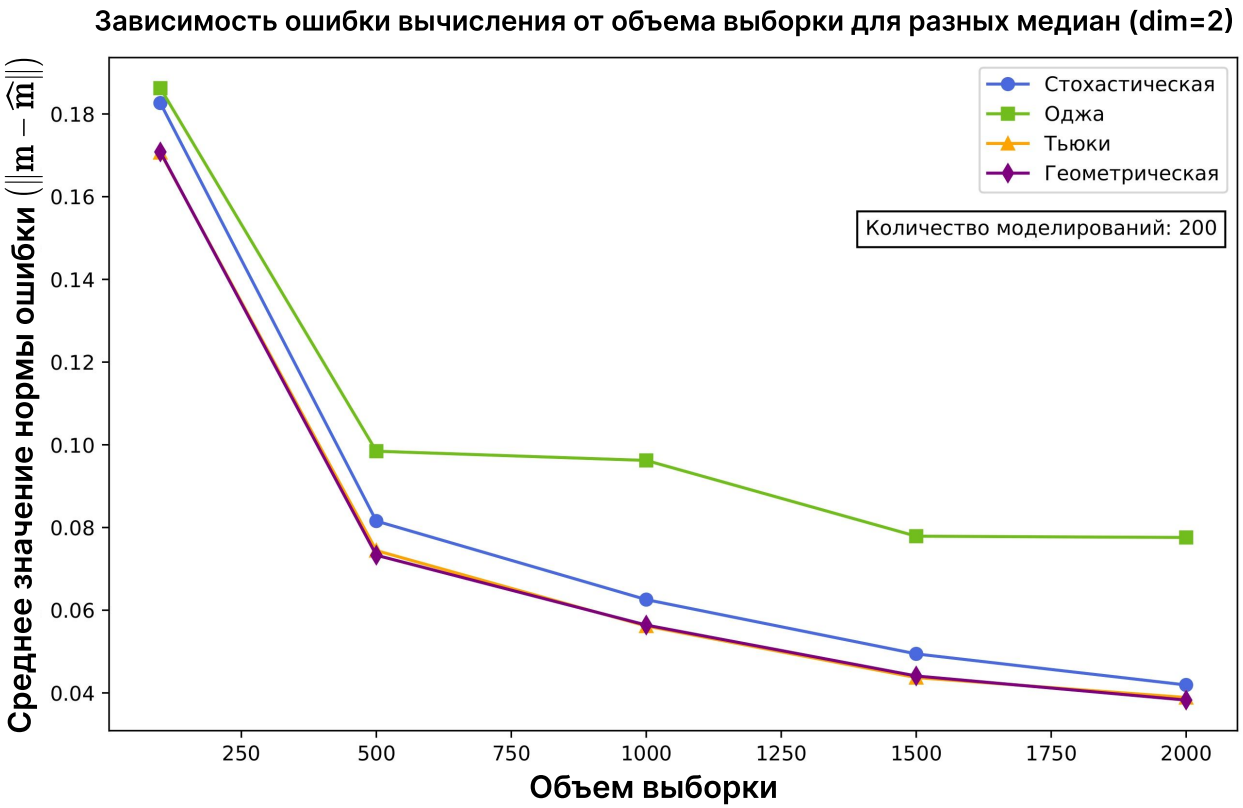
Исследование робастности

Теперь фиксируем количество итераций вместо точности и на тех же данных запустим алгоритм.



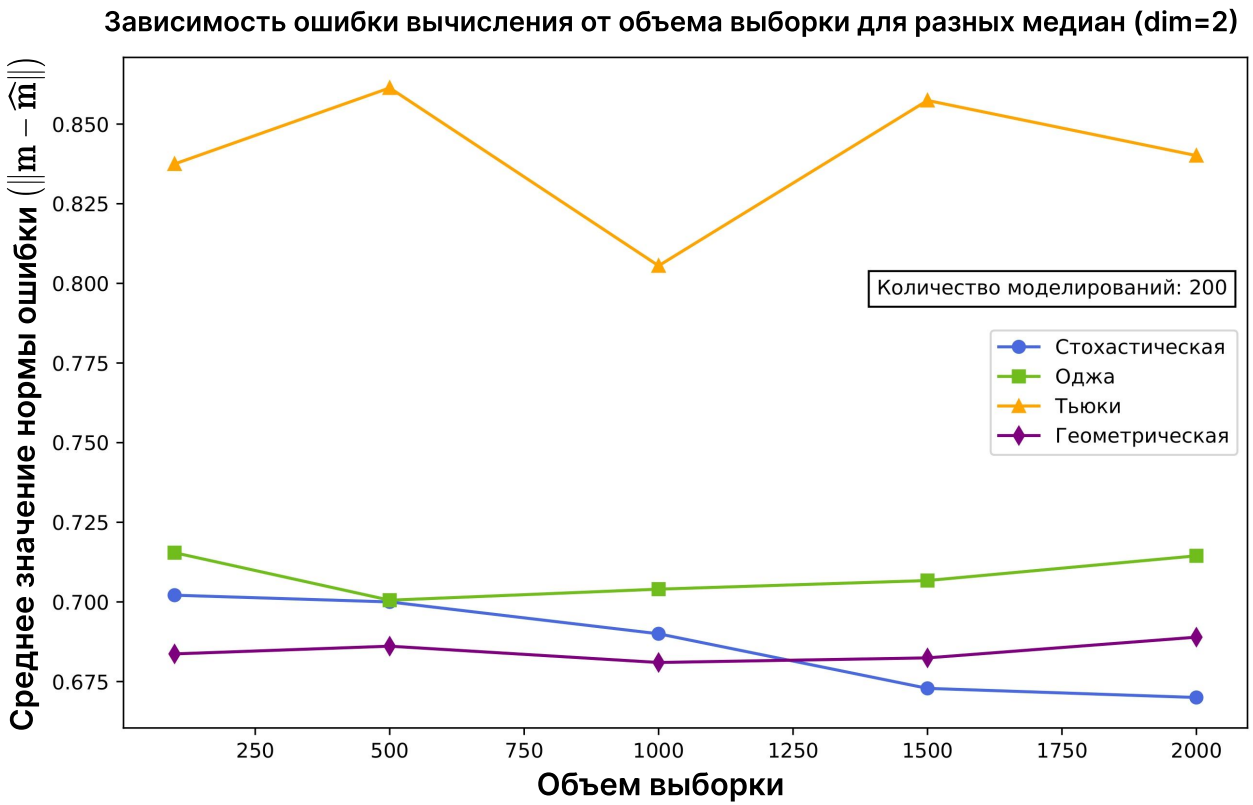
Сравнение результатов работы с другими алгоритмами для нормального распределения

Рассмотрим изменение ошибки для разных объемов выборки в случае нормального распределения с математическим ожиданием равным $(0,\ 0)^{\mathbf{T}}$ и ковариационной матрицей $\mathrm{diag}\{1,\ 2\}$.



Сравнение результатов работы с другими алгоритмами для нормального распределения

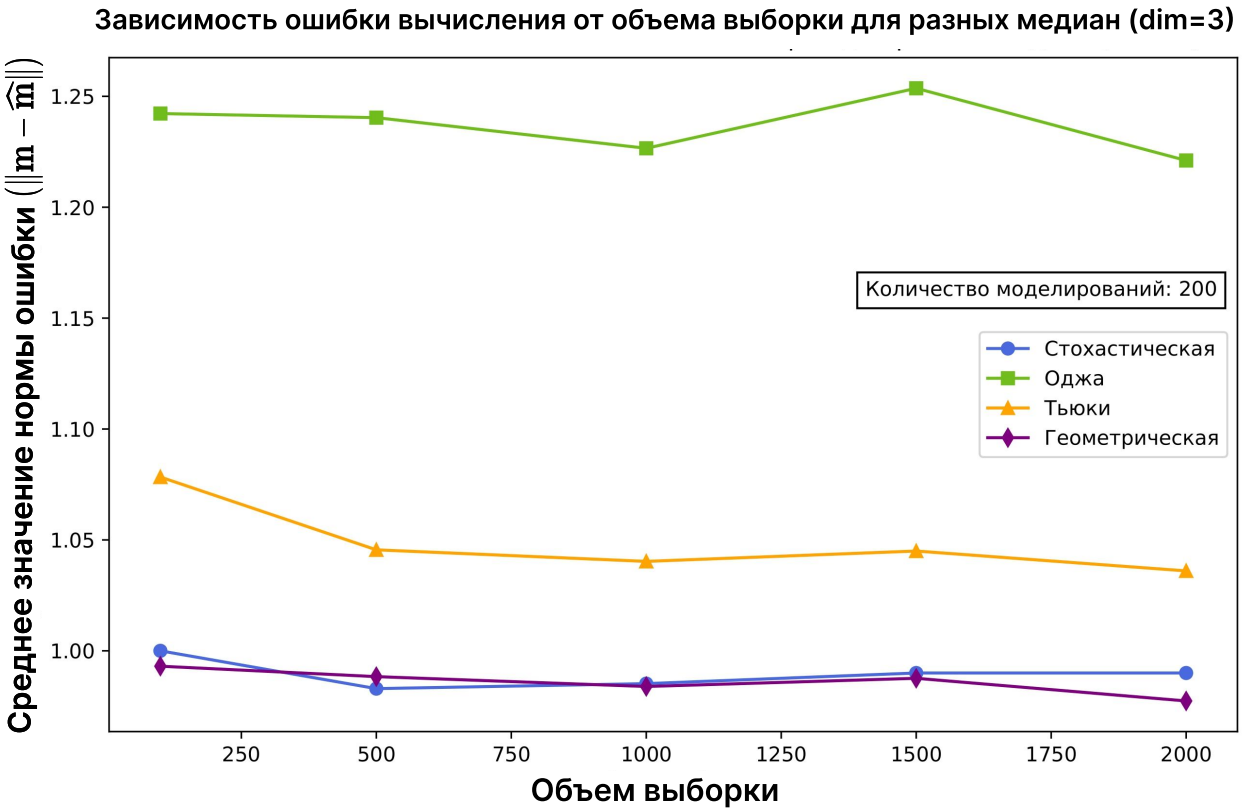
В выборке заменим 25% наблюдений на точку-выброс $(9,\ 12)^{\mathbf{T}}$ и посмотрим, как в данном случае изменяется ошибка.



Сравнение результатов работы с другими алгоритмами для нормального распределения

Пусть теперь математическое ожидание равно $(0,\ 0,\ 0)^{\mathbf{T}}$ и ковариационная матрица $\mathrm{diag}\{1,\ 2,\ 3\}$.

В выборке заменим 25% наблюдений на точку-выброс $(2,\ 5,\ 14)^{\mathbf{T}}$.



Другие результаты

Также мной были получены следующие результаты:

1. **Зависимость нормы ошибки от объема выборки**
   - Ошибка уменьшается пропорционально $1/\sqrt{n}$ с ростом объема выборки n.
2. **Зависимость количества итераций от точности**
   - Количество итераций k логарифмически растет с увеличением точности ε, что согласуется с полученным теоретическим результатом.
3. **Разброс оценки**
   - С увеличением объема выборки n дисперсия оценки уменьшается.

Заключение

1. Реализован новый алгоритм робастного оценивания параметра положения, исследована на модельном примере и строго доказана его скорость сходимости, а также временная сложность.
2. Проведено сравнение с распространёнными алгоритмами, которое показало, что предложенный алгоритм дает сопоставимые результаты и не уступает по эффективности.
3. Открытыми остались вопросы относительно проведения более глубокого исследования алгоритма:
   - распространении алгоритма для более сложных вероятностных распределений,
   - построении модификации алгоритма для работы с большими данными.