A Note on Efficient Compution of Rank Order Statistics

Daniel Grose
Lancaster University
dan.grose@lancaster.ac.uk

September 17, 2025

Abstract

The computation of both order and rank statistics is often necessary when using non-parametric methods in data analysis. However, it is not uncommon for the computational complexity associated with their calculation to be overlooked, both in terms of algorithm design, and data structure choice. Interestingly, this seems to be more often the case when there is a requirment that the data is frequently updated, as is the case, for example, when working data streams.

This note considers the cases of computing the Median Absolute Deviation, (an order statistic), and the Mann-Whitney U, (a rank statistic). Importantly, by using an appropriate data structure, it is shown that, for a data store of size n, the computational cost of an update is $\mathcal{O}(\log^2 n)$ for the Median Absolute Deviation, and $\mathcal{O}(\log n)$ for the Mann-Whitney U statistic .

1 Introduction

The median absolute deviation (MAD) is a widely used measure of statistical dispersion, known for its robustness against outliers. Computing the MAD of an arbitrary vector of size n requires $O(n)$ time in the best case, either by explicitly sorting the deviations or by applying a linear-time selection algorithm. In this note, we show that if the input vector is already sorted, the MAD can be computed in $O(\log n)$ time. The key observation is that the deviations from the median form two monotone subsequences, reducing the problem to finding the median of two sorted arrays, which is solvable in logarithmic time. This result has implications for both theoretical computer science and practical applications in robust statistics.

The median absolute deviation (MAD) is defined as

$$\mathrm{MAD}(x) = \mathrm{median}(|x_i - m|), \quad m = \mathrm{median}(x),$$

where $x = (x_1, \ldots, x_n)$ is a vector of real numbers. The MAD is widely used in statistics, machine learning, and data science as a robust measure of variability.

For an unsorted vector, computing the MAD requires $O(n)$ time: one must compute all deviations and find their median using either sorting ($O(n \log n)$) or a linear-time selection algorithm. The assumption of sorted input has not, to our knowledge, been fully exploited in prior work.

2 Preliminaries

Let $x = (x_1, \ldots, x_n)$ be a sorted vector, i.e., $x_1 \leq x_2 \leq \cdots \leq x_n$. Denote the median of x by m. Define the deviations as $d_i = |x_i - m|$ for $1 \leq i \leq n$.

We recall the well-studied problem of finding the median of two sorted arrays, which can be solved in $O(\log n)$ time using binary search partitioning.

3 Key Observation

Since x is sorted, the deviations split into two monotone subsequences:

- For $i <$ median index, $d_i = m - x_i$, which decreases as i decreases. Reversing the order yields an increasing sequence.
- For $i >$ median index, $d_i = x_i - m$, which increases as i increases.

Thus, the set of deviations is the union of two sorted sequences (plus a single 0 if n is odd).

4 Algorithm

1. Find the median m of x in $O(1)$ time.
2. Construct logical views of the left deviations $L = [m - x_k, m - x_{k-1}, \ldots, m - x_1]$ (which is sorted ascending) and the right deviations $R = [x_{k+1} - m, x_{k+2} - m, \ldots, x_n - m]$.
3. Use the median-of-two-sorted-arrays algorithm to compute the median of $L \cup R$ (plus 0 if n is odd).

5 Complexity Analysis

Finding the median of x is $O(1)$. Constructing L and R requires no materialization; indices can be handled implicitly. The median of two sorted arrays can be found in $O(\log n)$. Therefore, the total complexity is $O(\log n)$.

6 Discussion

This result improves upon the known $O(n)$ bound for MAD computation when the input is sorted. The savings are particularly relevant in scenarios where MAD must be computed repeatedly on pre-sorted data, such as streaming data pipelines, order statistics maintenance, and robust statistical preprocessing.

Future work includes exploring whether similar monotonicity arguments can be applied to accelerate computation of other robust statistics (e.g., interquartile range, trimmed means).

7 Conclusion

We have shown that the median absolute deviation of a sorted vector can be computed in logarithmic time by reducing the problem to finding the median of two sorted arrays. This observation bridges a gap between robust statistics and efficient selection algorithms.