Outlier Detection Under Interval Uncertainty: Algorithmic Solvability and Computational Complexity

Vladik Kreinovich[1], Luc Longpré[1], Praveen Patangay[1]
Scott Ferson[2], and Lev Ginzburg[2]

[1]Computer Science Department
University of Texas at El Paso
El Paso, TX 79968, USA
{vladik,longpre,praveen}@cs.utep.edu

[2]Applied Biomathematics
100 North Country Road
Setauket, NY 11733, USA
{scott,lev}@ramas.com

Abstract

In many application areas, it is important to detect outliers. Traditional engineering approach to outlier detection is that we start with some "normal" values $x_1, \ldots, x_n$, compute the sample average E, the sample standard variation σ, and then mark a value x as an outlier if x is outside the k_0-sigma interval $[E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$ (for some pre-selected parameter k_0). In real life, we often have only interval ranges $[\underline{x}_i, \overline{x}_i]$ for the normal values $x_1, \ldots, x_n$. In this case, we only have intervals of possible values for the bounds $E - k_0 \cdot \sigma$ and $E + k_0 \cdot \sigma$. We can therefore identify outliers as values that are outside all k_0-sigma intervals.

Once we identify a value as an outlier for a fixed k_0, it is also desirable to find out to what degree this value is an outlier, i.e., what is the largest value k_0 for which this value is an outlier.

In this paper, we analyze the computational complexity of these outlier detection problems, provide efficient algorithms that solve some of these problems (under reasonable conditions), and list related open problems.

1 Introduction

Outlier detection is important. In many application areas, it is important to detect *outliers*, i.e., unusual, abnormal values. In medicine, unusual values may indicate disease (see, e.g., [7, 19, 20]); in geophysics, abnormal values may indicate a mineral deposit or an erroneous measurement result (see, e.g., [5, 11, 15, 18]); in structural integrity testing, abnormal values may indicate faults in a structure (see, e.g., [2, 6, 7, 12, 13, 19, 20, 21]), etc.

Traditional engineering approach to outlier detection (see, e.g., [1, 14, 17]) is as follows:

- first, we collect measurement results $x_1, \ldots, x_n$ corresponding to normal situations;
- then, we compute the sample average $E \stackrel{\text{def}}{=} \dfrac{x_1 + \ldots + x_n}{n}$ of these normal values and the (sample) standard deviation $\sigma = \sqrt{V}$, where $V \stackrel{\text{def}}{=} \dfrac{(x_1 - E)^2 + \ldots + (x_n - E)^2}{n}$;
- finally, a new measurement result x is classified as an outlier if it is outside the interval $[L, U]$ (i.e., if either $x < L$ or $x > U$), where $L \stackrel{\text{def}}{=} E - k_0 \cdot \sigma$, $U \stackrel{\text{def}}{=} E + k_0 \cdot \sigma$, and $k_0 > 1$ is some pre-selected value (most frequently, $k_0 = 2$, 3, or 6).

Outlier detection under interval uncertainty. In some practical situations, we only have intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values of x_i. This happens, for example, if instead of observing the actual value x_i of the random variable, we observe the value $\widetilde{x}_i$ measured by an instrument with a known upper bound Δ_i on the measurement error; then, the actual (unknown) value is within the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. For different values $x_i \in \mathbf{x}_i$, we get different bounds L and U. Possible values of L form an interval – we will denote it by $\mathbf{L} \stackrel{\text{def}}{=} [\underline{L}, \overline{L}]$; possible values of U form an interval $\mathbf{U} = [\underline{U}, \overline{U}]$.

How do we now detect outliers? There are two possible approaches to this question: we can detect *possible* outliers and we can detect *guaranteed* outliers:

- a value x is a possible outlier if it is located outside one of the possible k_0-sigma intervals $[L, U]$ (but is may be inside some other possible interval $[L, U]$);
- a value x is a guaranteed outlier if it is located outside all possible k_0-sigma intervals $[L, U]$.

Which approach is more reasonable depends on a possible situation:

- if our main objective is not to miss an outlier, e.g., in structural integrity tests, when we do not want to risk launching a spaceship with a faulty part, it is reasonable to look for possible outliers;

- if we want to make sure that the value x is an outlier, e.g., if we are planning a surgery and we want to make sure that there is a micro-calcification before we start cutting the patient, then we would rather look for guaranteed outliers.

The two approaches can be described in terms of the endpoints of the intervals $\mathbf{L}$ and $\mathbf{U}$:

A value x guaranteed to be normal – i.e., it is not a possible outlier – if x belongs to the *intersection* of all possible intervals $[L, U]$; the intersection corresponds to the case when L is the largest and U is the smallest, i.e., this intersection is the interval $[\overline{L}, \underline{U}]$. So, if $x > \underline{U}$ or $x < \overline{L}$, then x is a possible outlier, else it is guaranteed to be a normal value.

If a value x is inside *one* of the possible intervals $[L, U]$, then it can still be normal; the only case when we are sure that the value x is an outlier is when x is outside *all* possible intervals $[L, U]$, i.e., is the value x does not belong to the *union* of all possible intervals $[L, U]$ of normal values; this union is equal to the interval $[\underline{L}, \overline{U}]$. So, if $x > \overline{U}$ or $x < \underline{L}$, then x is a guaranteed outlier, else it can be a normal value.

In real life, the situation may be slightly more complicated because, as we have mentioned, measurements often come with interval inaccuracy; so, instead of the exact value x of the measured quantity, we get an interval $\mathbf{x} = [\underline{x}, \overline{x}]$ of possible values of this quantity.

In this case, we have a slightly more complex criterion for outlier detection:

- the actual (unknown) value of the measured quantity is a possible outlier if some value x from the interval $[\underline{x}, \overline{x}]$ is a possible outlier, i.e., is outside the intersection $[\overline{L}, \underline{U}]$; thus, the value is a possible outlier if one of the two inequalities hold: $\underline{x} < \overline{L}$ or $\underline{U} < \overline{x}$.
- the actual (unknown) value of the measured quantity is guaranteed to be an outlier if all possible values x from the interval $[\underline{x}, \overline{x}]$ are guaranteed to be outliers (i.e., are outside the union $[\underline{L}, \overline{U}]$); thus, the value is a guaranteed outlier if one of the two inequalities hold: $\overline{x} < \underline{L}$ or $\overline{U} < \underline{x}$.

Thus:

- to detect possible outliers, we must be able to compute the values $\overline{L}$ and $\underline{U}$;
- to detect guaranteed outliers, we must be able to compute the values $\underline{L}$ and $\overline{U}$.

In this paper, we consider the problem of computing these bounds.

Once we identify a value as an outlier for a fixed k_0, it is also desirable to find out to what degree this value is an outlier, i.e., what is the largest value k_0 for which this value is an outlier. In this paper, we analyze the algorithmic solvability and computational complexity of this problem as well.

Some of the results from this paper have been announced in [9, 10].

What was known before. As we discussed in the introduction, to detect outliers under interval uncertainty, we must be able to compute the range $\mathbf{L} = [\underline{L}, \overline{L}]$ of possible values of $L = E - k_0 \cdot \sigma$ and the range $\mathbf{U} = [\underline{U}, \overline{U}]$ of possible values of $U = E + k_0 \cdot \sigma$.

In [3, 4], we have shown how to compute the intervals $\mathbf{E} = [\underline{E}, \overline{E}]$ and $[\underline{\sigma}, \overline{\sigma}]$ of possible values for E and σ. In principle, we can use the general ideas of interval computations to combine these intervals and conclude, e.g., that L always belongs to the interval $\mathbf{E} - k_0 \cdot [\underline{\sigma}, \overline{\sigma}]$. However, as often happens in interval computations, the resulting interval for L is *wider* than the actual range – wider because the values E and σ are computed based on the same inputs $x_1, \ldots, x_n$ and cannot, therefore, change independently.

We mark a value x as an outlier if it is outside the interval $[L, U]$. Thus, if, instead of the actual ranges for L and U, we use wider intervals, we may miss some outliers. It is therefore important to compute the *exact* ranges for L and U. In this paper, we show how to compute these exact ranges.

2 Detecting Possible Outliers

To find possible outliers, we must know the values $\underline{U}$ and $\overline{L}$. In this section, we design *feasible* algorithms for computing the exact lower bound $\underline{U}$ of the function U and the exact upper bound $\overline{L}$ of the function L. Specifically, our algorithms are *quadratic-time*, i.e., require $O(n^2)$ computational steps (arithmetic operations or comparisons) for n interval data points $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.

The algorithm $\underline{\mathcal{A}}_U$ for computing $\underline{U}$ is as follows:

- First, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}$; take $x_{(0)} = -\infty$ and $x_{(2n+1)} = +\infty$. Thus, the real line is divided into $2n+1$ zones $(x_{(0)}, x_{(1)}]$, $[x_{(1)}, x_{(2)}]$, $\ldots$, $[x_{(2n-1)}, x_{(2n)}]$, $[x_{(2n)}, x_{(2n+1)})$.
- For each of these zones $[x_{(k)}, x_{(k+1)}]$, $k = 0, 1, \ldots, 2n$, we compute the values
$$e_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} \underline{x}_i + \sum_{j:\overline{x}_j \le x_{(k)}} \overline{x}_j,$$
$$m_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} (\underline{x}_i)^2 + \sum_{j:\overline{x}_j \le x_{(k)}} (\overline{x}_j)^2,$$
and $n_k =$ the total number of such i's and j's. Then, we solve the quadratic equation
$$A - B \cdot \mu + C \cdot \mu^2 = 0,$$
where
$$A \stackrel{\text{def}}{=} e_k^2 \cdot (1 + \alpha^2) - \alpha^2 \cdot m_k \cdot n; \quad \alpha \stackrel{\text{def}}{=} 1/k_0,$$
$$B \stackrel{\text{def}}{=} 2 \cdot e_k \cdot \left((1 + \alpha^2) \cdot n_k - \alpha^2 \cdot n\right); \quad C \stackrel{\text{def}}{=} n_k \cdot \left((1 + \alpha^2) \cdot n_k - \alpha^2 \cdot n\right).$$

We consider only those solutions for which $\mu \cdot n_k \le e_k$ and $\mu \in [x_{(k)}, x_{(k+1)}]$. For each such solution, we compute the values of

$$E_k = \frac{e_k}{n} + \frac{n - n_k}{n} \cdot \mu, \quad M_k = \frac{m_k}{n} + \frac{n - n_k}{n} \cdot \mu^2,$$

and $U_k = E_k + k_0 \cdot \sqrt{M_k - (E_k)^2}$.

- Finally, we return the smallest of the values U_k as $\underline{U}$.

Theorem 2.1. *The algorithm $\underline{A}_U$ always computes $\underline{U}$ in quadratic time.*

Example. Let us illustrate this algorithm on a simple example when $k_0 = 2$, and we have $n = 2$ intervals $\mathbf{x}_1 = [-2, -1]$ and $\mathbf{x}_2 = [1, 2]$. In this case, sorting leads to $x_{(1)} = -2$, $x_{(2)} = -1$, $x_{(3)} = 1$, and $x_{(4)} = 2$. Thus, the real line is divided into 5 zones: $(-\infty, -2]$, $[-2, -1]$, $[-1, 1]$, $[1, 2]$, and $[2, +\infty)$. Let us show the computations for each of these zones:

- For the first zone $(x_{(0)}, x_{(1)}] = (-\infty, -2]$, we have $\underline{x}_i \ge -2$ for all i, hence $e_0 = (-2) + 1 = -1$, $m_0 = (-2)^2 + 1^2 = 5$, and $n_0 = 2$. Here, $\alpha = 1/k_0 = 1/2$, hence

$$A = 1^2 \cdot \left(1 + \frac{1}{4}\right) - \frac{1}{4} \cdot 5 \cdot 2 = \frac{5}{4} - \frac{5}{2} = -\frac{5}{4},$$

$$B = 2 \cdot (-1) \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 2 - \frac{1}{4} \cdot 2\right) = -4;$$

$$C = 5 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 2 - \frac{1}{4} \cdot 2\right) = 10.$$

The corresponding equation $10 \cdot \mu^2 + 4 \cdot \mu - 5/4 = 0$ has two roots

$$\mu_{1,2} = \frac{-4 \pm \sqrt{4^2 + 4 \cdot (5/4) \cdot 10}}{2 \cdot 10} = \frac{-4 \pm \sqrt{66}}{20}.$$

None of these roots is in the zone $(-\infty, -2]$.

- For the second zone $[-2, -1]$, we have $e_1 = 1$, $m_1 = 1^2 = 1$, and $n_1 = 1$. In this case,

$$A = 1 \cdot \left(1 + \frac{1}{4}\right) - \frac{1}{4} \cdot 1 \cdot 2 = \frac{3}{4};$$

$$B = 2 \cdot 1 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 1 - \frac{1}{4} \cdot 2\right) = 1.5;$$

$$C = 1 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 1 - \frac{1}{4} \cdot 2\right) = 0.75.$$

If we divide both sides of the corresponding equation $0.75 \cdot \mu^2 - 1.5 \cdot \mu + 0.75 = 0$ by 0.75, we get the equation $\mu^2 - 2\mu + 1 = 0$, which has the double root $\mu_{1,2} = 1$ that is outside the zone.

- For the third zone $[-1, 1]$, we have $e_2 = (-1)+1 = 0$, $m_2 = (-1)^2+1^2 = 2$, and $n_2 = 2$. Here, $A = -(1/4) \cdot 2 \cdot 2 = -1$, $B = 0$, and

  $$C = 2 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 2 - \frac{1}{4} \cdot 2\right) = 4.$$

  The corresponding quadratic equation is $4\mu^2 - 1 = 0$, hence $\mu^2 = 1/4$ and $\mu_{1,2} = \pm 1/2$. Both roots are within the zone. Out of these two roots, only $\mu = -0.5$ satisfies the inequality $\mu \cdot n_2 \le e_2 = 0$. For this root,

  $$E_2 = \frac{0}{2} + \frac{2-2}{2} \cdot (-0.5) = 0;$$

  $$M_2 = \frac{2}{2} + \frac{2-2}{2} \cdot (-0.5)^2 = 1,$$

  and

  $$U_2 = 0 + 2 \cdot \sqrt{1 - 0^2} = 2.$$

- For the fourth zone $[1, 2]$, we have $e_3 = -1$, $m_3 = (-1)^2 = 1$, $n_3 = 1$, hence

  $$A = 1 \cdot \left(1 + \frac{1}{4}\right) - \frac{1}{4} \cdot 1 \cdot 2 = \frac{3}{4};$$

  $$B = 2 \cdot (-1) \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 1 - \frac{1}{4} \cdot 2\right) = -1.5;$$

  $$C = 1 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 1 - \frac{1}{4} \cdot 2\right) = 0.75.$$

  If we divide both sides of the corresponding equation $0.75 \cdot \mu^2 + 1.5 \cdot \mu + 0.75 = 0$ by 0.75, we get the equation $\mu^2 + 2\mu + 1 = 0$, which has the double root $\mu_{1,2} = -1$ that is outside the zone.

- For the fifth zone $[2, \infty)$, we have $\overline{x}_i \le 2$ for all i, hence $e_4 = (-1)+2 = 1$, $m_4 = (-1)^2 + 2^2 = 5$, and $n_0 = 2$. Here,

  $$A = 1^2 \cdot \left(1 + \frac{1}{4}\right) - \frac{1}{4} \cdot 5 \cdot 2 = \frac{5}{4} - \frac{5}{2} = -\frac{5}{4},$$

  $$B = 2 \cdot 1 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 2 - \frac{1}{4} \cdot 2\right) = 4;$$

  $$C = 5 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 2 - \frac{1}{4} \cdot 2\right) = 10.$$

  The corresponding equation $10 \cdot \mu^2 - 4 \cdot \mu - 5/4 = 0$ has two roots

  $$\mu_{1,2} = \frac{4 \pm \sqrt{4^2 + 4 \cdot (5/4) \cdot 10}}{2 \cdot 10} = \frac{4 \pm \sqrt{66}}{20}.$$

  None of these roots is in the zone $[2, +\infty)$.

The only value U_k is 2, hence $\underline{U} = 2$.

The algorithm $\overline{\mathcal{A}}_L$ for computing $\overline{L}$ is as follows:

- First, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}$; take $x_{(0)} = -\infty$ and $x_{(2n+1)} = +\infty$. Thus, the real line is divided into $2n+1$ zones $(x_{(0)}, x_{(1)}]$, $[x_{(1)}, x_{(2)}]$, $\ldots$, $[x_{(2n-1)}, x_{(2n)}]$, $[x_{(2n)}, x_{(2n+1)})$.
- For each of these zones $[x_{(k)}, x_{(k+1)}]$, $k = 0, 1, \ldots, 2n$, we compute the values
$$e_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} \underline{x}_i + \sum_{j:\overline{x}_j \le x_{(k)}} \overline{x}_j,$$
$$m_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} (\underline{x}_i)^2 + \sum_{j:\overline{x}_j \le x_{(k)}} (\overline{x}_j)^2,$$
and $n_k =$ the total number of such i's and j's. Then, we solve the quadratic equation
$$A - B \cdot \mu + C \cdot \mu^2 = 0,$$
where
$$A \stackrel{\text{def}}{=} e_k^2 \cdot (1 + \alpha^2) - \alpha^2 \cdot m_k \cdot n; \quad \alpha \stackrel{\text{def}}{=} 1/k_0,$$
$$B \stackrel{\text{def}}{=} 2 \cdot e_k \cdot \left((1+\alpha^2) \cdot n_k - \alpha^2 \cdot n\right); \quad C \stackrel{\text{def}}{=} n_k \cdot \left((1+\alpha^2) \cdot n_k - \alpha^2 \cdot n\right).$$
We consider only those solutions for which $\mu \cdot n_k \ge e_k$ and $\mu \in [x_{(k)}, x_{(k+1)}]$. For each such solution, we compute the values of
$$E_k = \frac{e_k}{n} + \frac{n - n_k}{n} \cdot \mu, \quad M_k = \frac{m_k}{n} + \frac{n - n_k}{n} \cdot \mu^2,$$
and $L_k = E_k - k_0 \cdot \sqrt{M_k - (E_k)^2}$.
- Finally, we return the largest of the values L_k as $\overline{L}$.

Theorem 2.2. *The algorithm $\overline{\mathcal{A}}_L$ always computes $\overline{L}$ in quadratic time.*

Example. Let us illustrate this algorithm on the same simple example when $k_0 = 2$, and we have $n = 2$ intervals $\mathbf{x}_1 = [-2, -1]$ and $\mathbf{x}_2 = [1, 2]$. In this case, sorting leads to $x_{(1)} = -2$, $x_{(2)} = -1$, $x_{(3)} = 1$, and $x_{(4)} = 2$. Thus, the real line is divided into 5 zones: $(-\infty, -2]$, $[-2, -1]$, $[-1, 1]$, $[1, 2]$, and $[2, +\infty)$. Let us show the computations for each of these zones:

- For the first zone $(x_{(0)}, x_{(1)}] = (-\infty, -2]$, we have $\underline{x}_i \ge -2$ for all i, hence $e_0 = (-2) + 1 = -1$, $m_0 = (-2)^2 + 1^2 = 5$, and $n_0 = 2$. Here, $\alpha = 1/k_0 = 1/2$, hence
$$A = 1^2 \cdot \left(1 + \frac{1}{4}\right) - \frac{1}{4} \cdot 5 \cdot 2 = \frac{5}{4} - \frac{5}{2} = -\frac{5}{4},$$

$$B = 2 \cdot (-1) \cdot \left(\left(1 + \frac{1}{4} \right) \cdot 2 - \frac{1}{4} \cdot 2 \right) = -4;$$

$$C = 5 \cdot \left(\left(1 + \frac{1}{4} \right) \cdot 2 - \frac{1}{4} \cdot 2 \right) = 10.$$

The corresponding equation $10 \cdot \mu^2 + 4 \cdot \mu - 5/4 = 0$ has two roots

$$\mu_{1,2} = \frac{-4 \pm \sqrt{4^2 + 4 \cdot (5/4) \cdot 10}}{2 \cdot 10} = \frac{-4 \pm \sqrt{66}}{20}.$$

None of these roots is in the zone $(-\infty, -2]$.

- For the second zone $[-2, -1]$, we have $e_1 = 1$, $m_1 = 1^2 = 1$, and $n_1 = 1$. In this case,

$$A = 1 \cdot \left(1 + \frac{1}{4} \right) - \frac{1}{4} \cdot 1 \cdot 2 = \frac{3}{4};$$

$$B = 2 \cdot 1 \cdot \left(\left(1 + \frac{1}{4} \right) \cdot 1 - \frac{1}{4} \cdot 2 \right) = 1.5;$$

$$C = 1 \cdot \left(\left(1 + \frac{1}{4} \right) \cdot 1 - \frac{1}{4} \cdot 2 \right) = 0.75.$$

If we divide both sides of the corresponding equation $0.75 \cdot \mu^2 - 1.5 \cdot \mu + 0.75 = 0$ by 0.75, we get the equation $\mu^2 - 2\mu + 1 = 0$, which has the double root $\mu_{1,2} = 1$ that is outside the zone.

- For the third zone $[-1, 1]$, we have $e_2 = (-1) + 1 = 0$, $m_2 = (-1)^2 + 1^2 = 2$, and $n_2 = 2$. Here, $A = -(1/4) \cdot 2 \cdot 2 = -1$, $B = 0$, and

$$C = 2 \cdot \left(\left(1 + \frac{1}{4} \right) \cdot 2 - \frac{1}{4} \cdot 2 \right) = 4.$$

The corresponding quadratic equation is $4\mu^2 - 1 = 0$, hence $\mu^2 = 1/4$ and $\mu_{1,2} = \pm 1/2$. Both roots are within the zone. Out of these two roots, only $\mu = 0.5$ satisfies the inequality $\mu \cdot n_2 \geq e_2 = 0$. For this root,

$$E_2 = \frac{0}{2} + \frac{2-2}{2} \cdot 0.5 = 0;$$

$$M_2 = \frac{2}{2} + \frac{2-2}{2} \cdot 0.5^2 = 1,$$

and

$$L_2 = 0 - 2 \cdot \sqrt{1 - 0^2} = 2.$$

- For the fourth zone $[1,2]$, we have $e_3 = -1$, $m_3 = (-1)^2 = 1$, $n_3 = 1$, hence
$$A = 1 \cdot \left(1 + \frac{1}{4}\right) - \frac{1}{4} \cdot 1 \cdot 2 = \frac{3}{4};$$
$$B = 2 \cdot (-1) \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 1 - \frac{1}{4} \cdot 2\right) = -1.5;$$
$$C = 1 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 1 - \frac{1}{4} \cdot 2\right) = 0.75.$$
If we divide both sides of the corresponding equation $0.75 \cdot \mu^2 + 1.5 \cdot \mu + 0.75 = 0$ by 0.75, we get the equation $\mu^2 + 2\mu + 1 = 0$, which has the double root $\mu_{1,2} = -1$ that is outside the zone.

- For the fifth zone $[2, \infty)$, we have $\overline{x}_i \le 2$ for all i, hence $e_4 = (-1) + 2 = 1$, $m_4 = (-1)^2 + 2^2 = 5$, and $n_0 = 2$. Here,
$$A = 1^2 \cdot \left(1 + \frac{1}{4}\right) - \frac{1}{4} \cdot 5 \cdot 2 = \frac{5}{4} - \frac{5}{2} = -\frac{5}{4},$$
$$B = 2 \cdot 1 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 2 - \frac{1}{4} \cdot 2\right) = 4;$$
$$C = 5 \cdot \left(\left(1 + \frac{1}{4}\right) \cdot 2 - \frac{1}{4} \cdot 2\right) = 10.$$
The corresponding equation $10 \cdot \mu^2 - 4 \cdot \mu - 5/4 = 0$ has two roots
$$\mu_{1,2} = \frac{4 \pm \sqrt{4^2 + 4 \cdot (5/4) \cdot 10}}{2 \cdot 10} = \frac{4 \pm \sqrt{66}}{20}.$$
None of these roots is in the zone $[2, +\infty)$.

The only value L_k is -2, hence $\overline{L} = -2$.

3 In General, Detecting Guaranteed Outliers is NP-Hard

As we have mentioned in Section 1, to be able to detect guaranteed outliers, we must be able to compute the values $\underline{L}$ and $\overline{U}$. In general, this is an NP-hard problem:

Theorem 3.1. *For every $k_0 > 1$, computing the upper endpoint $\overline{U}$ of the interval $[\underline{U}, \overline{U}]$ of possible values of $U = E + k_0 \cdot \sigma$ is NP-hard.*

Theorem 3.2. *For every $k_0 > 1$, computing the lower endpoint $\underline{L}$ of the interval $[\underline{L}, \overline{L}]$ of possible values of $L = E - k_0 \cdot \sigma$ is NP-hard.*

(For readers' convenience, all the proofs are placed in the special Proofs section).

Comment. For interval data, the NP-hardness of computing the upper bound for σ was proven in [3, 4]. The general overview of NP-hardness of computational problems in interval context is given in [8].

4 How Can We Actually Detect Guaranteed Outliers?

How can we actually compute these values? First, we will show that if $1 + (1/k_0)^2 < n$ (which is true, e.g., if $k_0 > 1$ and $n \geq 2$), then the maximum of U (correspondingly, the minimum of L) is always attained at some combination of endpoints of the intervals $\mathbf{x}_i$; thus, in principle, to determine the values $\overline{U}$ and $\underline{L}$, it is sufficient to try all 2^n combinations of values $\underline{x}_i$ and $\overline{x}_i$:

Theorem 4.1. *If $1 + (1/k_0)^2 < n$, then the maximum of the function U on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$ is attained at one of its vertices, i.e., when for every i, either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$.*

Theorem 4.2. *If $1 + (1/k_0)^2 < n$, then the minimum of the function L on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$ is attained at one of its vertices, i.e., when for every i, either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$.*

NP-hard means, crudely speaking, that there are no general ways for solving all particular cases of this problem (i.e., computing $\overline{V}$) in reasonable time.

However, we show that there are algorithms for computing $\overline{U}$ and $\underline{L}$ for many reasonable situations. Namely, we propose efficient algorithms that compute $\overline{U}$ and $\underline{L}$ for the case when all the interval midpoints ("measured values") $\widetilde{x}_i \overset{\text{def}}{=} (\underline{x}_i + \overline{x}_i)/2$ are definitely different from each other, in the sense that the "narrowed" intervals

$$\left[\widetilde{x}_i - \frac{1+\alpha^2}{n} \cdot \Delta_i, \widetilde{x}_i + \frac{1+\alpha^2}{n} \cdot \Delta_i\right]$$

– where $\alpha = 1/k_0$ and $\Delta_i \overset{\text{def}}{=} (\underline{x}_i - \overline{x}_i)/2$ is the interval's half-width – do not intersect with each other.

The algorithm $\overline{\mathcal{A}}_U$ is as follows:

- First, we sort all $2n$ endpoints of the narrowed intervals $\widetilde{x}_i - \frac{1+\alpha^2}{n} \cdot \Delta_i$ and $\widetilde{x}_i + \frac{1+\alpha^2}{n} \cdot \Delta_i$ into a sequence $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(2n)}$. This enables us to divide the real line into $2n+1$ segments ("small intervals") $[x_{(i)}, x_{(i+1)}]$, where we denoted $x_{(0)} \overset{\text{def}}{=} -\infty$ and $x_{(2n+1)} \overset{\text{def}}{=} +\infty$.

- For each of zones $[x_{(i)}, x_{(i+1)}]$, we do the following: for each j from 1 to n, we pick the following value of x_j:

  - if $x_{(i+1)} < \widetilde{x}_j - \dfrac{1+\alpha^2}{n} \cdot \Delta_j$, then we pick $x_j = \overline{x}_j$;
  - if $x_{(i+1)} > \widetilde{x}_j + \dfrac{1+\alpha^2}{n} \cdot \Delta_j$, then we pick $x_j = \underline{x}_j$;
  - for all other j, we consider both possible values $x_j = \overline{x}_j$ and $x_j = \underline{x}_j$.

  As a result, we get one or several sequences of x_j. For each of these sequences, we check whether, for the selected values $x_1, \ldots, x_n$, the value of $E - \alpha \cdot \sigma$ is indeed within this small interval, and if it is, compute the value $U = E + k_0 \cdot \sigma$.

- Finally, we return the largest of the computed values U as $\overline{U}$.

Theorem 4.3. *Let* $1+(1/k_0)^2 < n$. *The algorithm* $\overline{\mathcal{A}}_U$ *computes* $\overline{U}$ *in quadratic time for all the cases in which the "narrowed" intervals do not intersect with each other.*

A similar algorithm $\underline{\mathcal{A}}_L$ can be designed for computing $\underline{L}$:

- First, we sort all $2n$ endpoints of the narrowed intervals $\widetilde{x}_i - \dfrac{1+\alpha^2}{n} \cdot \Delta_i$ and $\widetilde{x}_i + \dfrac{1+\alpha^2}{n} \cdot \Delta_i$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}$. This enables us to divide the real line into $2n+2$ segments ("small intervals") $[x_{(i)}, x_{(i+1)}]$, where we denoted $x_{(0)} \stackrel{\text{def}}{=} -\infty$ and $x_{(2n+1)} \stackrel{\text{def}}{=} +\infty$.

- For each of the zones $[x_{(i)}, x_{(i+1)}]$, we do the following: for each j from 1 to n, we pick the following value of x_j:

  - if $x_{(i+1)} < \widetilde{x}_j - \dfrac{1+\alpha^2}{n} \cdot \Delta_j$, then we pick $x_j = \overline{x}_j$;
  - if $x_{(i+1)} > \widetilde{x}_j + \dfrac{1+\alpha^2}{n} \cdot \Delta_j$, then we pick $x_j = \underline{x}_j$;
  - for all other j, we consider both possible values $x_j = \overline{x}_j$ and $x_j = \underline{x}_j$.

  As a result, we get one or several sequences of x_j. For each of these sequences, we check whether, for the selected values $x_1, \ldots, x_n$, the value of $E + \alpha \cdot \sigma$ is indeed within this small interval, and if it is, compute the value $L = E - k_0 \cdot \sigma$.

- Finally, we return the smallest of the computed values L as $\underline{L}$.

Theorem 4.4. *Let* $1+(1/k_0)^2 < n$. *The algorithm* $\underline{\mathcal{A}}_L$ *compute* $\underline{L}$ *in quadratic time for all the cases in which the "narrowed" intervals do not intersect with each other.*

These algorithms also work when, for some fixed C, no more than C "narrowed" intervals can have a common point:

Theorem 4.5. *Let $1+(1/k_0)^2 < n$. For every positive integer C, the algorithm $\overline{\mathcal{A}}_U$ computes $\overline{U}$ in quadratic time for all the cases in which no more than C "narrowed" intervals can have a common point.*

Theorem 4.6. *Let $1+(1/k_0)^2 < n$. For every positive integer C, the algorithm $\underline{\mathcal{A}}_L$ computes $\underline{L}$ in quadratic time for all the cases in which no more than C "narrowed" intervals can have a common point.*

The corresponding computation times are quadratic in n but grow exponentially with C. So, when C grows, this algorithm requires more and more computation time. It is worth mentioning that the examples on which we prove NP-hardness (see proof of Theorem 3.1) correspond to the case when $n/2$ out of n narrowed intervals have a common point.

5 Computing Degree of Outlier-Ness

Formulation of the problem. As we mentioned in the Introduction, once we identify a value x as an outlier for a fixed k_0, it is also desirable to find out to what degree this value is an outlier, i.e., what is the largest value k_0 for which this value x is outside the corresponding k_0-sigma interval $[E-k_0\cdot\sigma, E+k_0\cdot\sigma]$.

If we know the exact values of the measurement results $x_1,\ldots,x_n$, then we can compute the exact values of E and σ and thus, determine this "degree of outlier-ness" as the ratio $r \stackrel{\rm def}{=} |x-E|/\sigma$. If we only know the intervals $\mathbf{x}_i$ of possible values of x_i, then different values $x_i \in \mathbf{x}_i$ may lead to different values of this ratio. In this situation, it is desirable to know the *interval* of possible values of r.

Simplifcation of the problem. In order to compute this interval, let us first reduce the problem of computing this interval to a simpler problem. This reduction will be done in three steps.

First, it turns out that the value of r does not change if, instead of the original variables x_i with values from intervals $\mathbf{x}_i$, we consider new variables $x'_i \stackrel{\rm def}{=} x_i - x$ and a new value $x'=0$. Indeed, in this case, $E'=E-x$ hence $E'-x'=E-x$, and the standard deviation σ does not change if we simply shift all the values x_i. Thus, without losing generality, we can assume that $x=0$, and we are therefore interested in the ratio $|E|/\sigma$.

Second, the lower bound of the ratio r is attained when the reverse ratio $1/r=\sigma/|E|$ is the largest, and vice versa. Thus, to find the interval of possible values for $|E|/\sigma$, it is sufficient to find the interval of possible values of $\sigma/|E|$. Computing this interval is, in its turn, equivalent to computing the interval for the square V/E^2 of the reverse ratio $1/r$.

Finally, since $V = M - E^2$, where $M \stackrel{\rm def}{=} \dfrac{x_1^2 + \ldots + x_n^2}{n}$ is the second moment, we have $V/E^2 = M/E^2 - 1$, so computing the bounds for V/E^2 is equivalent to computing the bounds for the ratio $R \stackrel{\rm def}{=} M/E^2$.

In this section, we will describe how to compute the bounds $\underline{R}$ and $\overline{R}$ for the ratio R; based on these bounds, we can compute the desired bounds on k_0.

Computing $\underline{R}$: algorithm. The algorithm $\underline{\mathcal{A}}_R$ for computing $\underline{R}$ is as follows. If all the original intervals have a common point, then we take $\underline{R} \stackrel{\rm def}{=} 1$. Otherwise, we do the following:

- First, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}$; take $x_{(0)} = -\infty$ and $x_{(2n+1)} = +\infty$. Thus, the real line is divided into $2n+1$ zones $(x_{(0)}, x_{(1)}]$, $[x_{(1)}, x_{(2)}]$, $\ldots$, $[x_{(2n-1)}, x_{(2n)}]$, $[x_{(2n)}, x_{(2n+1)})$.
- For each of these zones $[x_{(k)}, x_{(k+1)}]$, $k = 0, 1, \ldots, 2n$, we compute the values
$$e_k \stackrel{\rm def}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} \underline{x}_i + \sum_{j:\overline{x}_j \le x_{(k)}} \overline{x}_j,$$
$$m_k \stackrel{\rm def}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} (\underline{x}_i)^2 + \sum_{j:\overline{x}_j \le x_{(k)}} (\overline{x}_j)^2,$$
and $n_k =$ the total number of such i's and j's. Then, we find $\lambda_k \stackrel{\rm def}{=} m_k/e_k$. If $\lambda_k \in [x_{(k)}, x_{(k+1)}]$, then we compute
$$E_k = \frac{e_k}{n} + \frac{n - n_k}{n} \cdot \lambda_k, \quad M_k = \frac{m_k}{n} + \frac{n - n_k}{n} \cdot \lambda_k^2,$$
and $R_k \stackrel{\rm def}{=} M_k/E_k^2$.
- Finally, we return the smallest of the values R_k as $\underline{R}$.

Theorem 5.1 *The algorithm $\underline{\mathcal{A}}_R$ always computes $\underline{R}$ in quadratic time.*

Computing $\underline{R}$: example. Let us trace this algorithm on a simple example when $n = 2$, $\mathbf{x}_1 = [1, 2]$, and $\mathbf{x}_2 = [3, 4]$. In this case, $x_{(1)} = 1$, $x_{(2)} = 2$, $x_{(3)} = 3$, $x_{(4)} = 4$, so we have 5 zones: $(-\infty, 1]$, $[1, 2]$, $[2, 3]$, $[3, 4]$, and $[4, +\infty)$. Let us run the algorithm for each zone:

- For $(-\infty, 1]$, we have $e_0 = 1 + 3 = 4$, $m_0 = 1^2 + 3^2 = 10$, $n_0 = 2$, hence $\lambda_0 = 10/4 = 2.5$ which is outside the zone.
- For $[1, 2]$, we have $e_1 = 3$, $m_1 = 3^2 = 9$, $n_1 = 1$, hence $\lambda_1 = 9/3 = 3$ which is outside the zone.

- For $[2,3]$, we have $e_2 = 3+2 = 5$, $m_2 = 3^2 + 3^2 = 13$, $n_2 = 2$, hence $\lambda_2 = 13/5 = 2.6$ which is within the zone. Here, $E_2 = 5/2$, $M_2 = 13/2$, hence
$$R_2 = \frac{13/2}{(5/2)^2} = \frac{13 \cdot 4}{25 \cdot 2} = \frac{52}{50} = 1.04.$$
- For $[3,4]$, we have $e_3 = 2$, $m_3 = 2^2 = 4$, $n_3 = 1$, so $\lambda_3 = 4/2 = 2$ which is outside the zone.
- For $[4,\infty)$, we have $e_4 = 2+4 = 6$, $m_4 = 2^2 + 4^2 = 20$, $n_4 = 4$, hence $\lambda_4 = 20/6 = 3.33\ldots$ is outside the zone.

The only value R_k is 1.04, so $\underline{R} = 1.04$.

The maximum $\overline{R}$ is always attained at the endpoints:

Theorem 5.2. *The maximum $\overline{R}$ of the function $R = M/E^2$ on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$ is attained at one of its vertices, i.e., when for every i, either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$.*

We are able to efficiently compute $\overline{R}$ if the "narrowed" intervals $[x_i^-, x_i^+]$ have few intersections, where:

$$x_i^- \stackrel{\text{def}}{=} \frac{\widetilde{x}_i}{1 + \dfrac{\Delta_i}{\underline{E} \cdot n}}; \quad x_i^+ \stackrel{\text{def}}{=} \frac{\widetilde{x}_i}{1 - \dfrac{\Delta_i}{\underline{E} \cdot n}}, \tag{1}$$

and $\underline{E} \stackrel{\text{def}}{=} \dfrac{\underline{x}_1 + \ldots + \underline{x}_n}{n}$ where $\widetilde{x}_i \stackrel{\text{def}}{=} (\underline{x}_i + \overline{x}_i)/2$ and $\Delta_i \stackrel{\text{def}}{=} (\underline{x}_i - \overline{x}_i)/2$.

The corresponding algorithm $\overline{\mathcal{A}}_R$ is as follows:

- First, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}$, take $x_{(0)} = -\infty$ and $x_{(2n+1)} = +\infty$, and thus divide the real line into $2n+1$ zones $(x_{(0)}, x_{(1)}]$, $[x_{(1)}, x_{(2)}]$, $\ldots$, $[x_{(2n-1)}, x_{(2n)}]$, $[x_{(2n)}, x_{(2n+1)})$.
- For each of these zones $[x_{(k)}, x_{(k+1)}]$, $k = 0, 1, \ldots, 2n$, and for each variable x_i, we take:
    - $x_i = \underline{x}_i$ if $x_i^+ \le x_{(k)}$;
    - $x_i = \overline{x}_i$ if $x_i^- \ge x_{(k+1)}$;
    - both values $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$ otherwise.

  For each of these combinations, we compute e_k, m_k, and $\lambda_k = m_k/e_k$, and check if λ_k is within the zone; if it is, we compute E_k, M_k, and $R_k = M_k/E_k^2$.

The largest of this computed value R_k is the desired upper endpoint $\overline{R}$.

Theorem 5.3. *For every positive integer C, the algorithm $\overline{\mathcal{A}}_R$ computes $\overline{R}$ in quadratic time for all the cases in which no more than C "narrowed" intervals can have a common point.*

6 Proofs

Proof of Theorem 2.1

1°. We will only prove Theorem 2.1; the proof of Theorem 2.2 is practically identical.

Let us first show that the algorithm $\underline{A}_U$ described in Section 2 is indeed correct.

Our proof of Theorem 2.1 is based on the fact that when the function $U(x_1,\ldots,x_n)$ attains its smallest possible value at some point $(x_1^{\rm opt},\ldots,x_n^{\rm opt})$, then, for every i, the corresponding function of one variable

$$U_i(x_i) \stackrel{\rm def}{=} U(x_1^{\rm opt},\ldots,x_{i-1}^{\rm opt},x_i,x_{i+1}^{\rm opt},\ldots,x_n^{\rm opt})$$

– the function that is obtained from $U(x_1,\ldots,x_n)$ by fixing the values of all the variables except for x_i – also attains its minimum at the value $x_i = x_i^{\rm opt}$.

A differentiable function of one variable attains its minimum on a closed interval either at one of its endpoints or at an internal point in which its first derivative is equal to 0.

By definition, $U = E + k_0 \cdot \sigma$. It is known that $\sigma = \sqrt{M - E^2}$, where $M \stackrel{\rm def}{=} (1/n) \cdot \sum\limits_{i=1}^n x_i^2$ is the sample second moment. Here,

$$\frac{\partial E}{\partial x_i} = \frac{1}{n}; \quad \frac{\partial M}{\partial x_i} = \frac{2 \cdot x_i}{n};$$

therefore,

$$\frac{\partial \sigma}{\partial x_i} = \frac{1}{2\sigma} \cdot \left(\frac{\partial M}{\partial x_i} - 2 \cdot E \cdot \frac{\partial E}{\partial x_i}\right) = \frac{1}{2\sigma} \cdot \left(\frac{2 \cdot x_i}{n} - 2 \cdot E \cdot \frac{1}{n}\right) = \frac{x_i - E}{\sigma \cdot n}.$$

Hence, we conclude that

$$\frac{dU_i}{dx_i} = \frac{\partial U}{\partial x_i} = \frac{1}{n} + k_0 \cdot \frac{x_i - E}{\sigma \cdot n}.$$

Therefore, this first derivative is equal to 0 when $\sigma + k_0 \cdot (x_i - E) = 0$, i.e., when $x_i = E - \alpha \cdot \sigma$, where $\alpha = 1/k_0$.

Thus, for the optimal values $x_1,\ldots,x_n$ for which U attains its minimum, for every i, we have either $x_i = \underline{x}_i$, or $x_i = \overline{x}_i$, or $x_i = E - \alpha \cdot \sigma$.

2°. Let us show that if the open interval $(\underline{x}_i, \overline{x}_i)$ contains the value $E - \alpha \cdot \sigma$, then the minimum of the function cannot be attained at points $\overline{x}_i$ or $\underline{x}_i$ and therefore, has to be attained at the value $x_i = E - \alpha \cdot \sigma$.

Let us show that the minimum cannot be attained for $x_i = \overline{x}_i$ (for $x_i = \underline{x}_i$, the proof is similar). We will prove this impossibility by reduction to a contradiction:

namely, we assume that the minimum is attained for $x_i = \overline{x}_i$, and we will deduce a contradiction from this assumption.

The fact that the minimum is attained for $x_i = \overline{x}_i$ means, in particular, that if we keep all the other values x_j the same but replace x_i by $x'_i = E - \alpha\cdot\sigma$, then the value U will not decrease. Let us denote the change in x_i by $\Delta x_i \stackrel{\rm def}{=} x_i - x'_i = \overline{x}_i - (E - \alpha\cdot\sigma)$; clearly, $\Delta x_i > 0$. We will denote the values of E, U, etc., that correspond to $(x_1,\ldots,x_{i-1},x'_i,x_{i+1},\ldots,x_n)$, by E', U', etc. In these terms, the desired inequality takes the form $U \le U'$, where $U = E + k_0\cdot\sqrt{M - E^2}$ and $U' = E' + k_0\cdot\sqrt{M' - (E')^2}$. It is convenient to multiply both sides of this inequality by $\alpha = 1/k_0$ and get an equivalent inequality $J \le J'$, where $J = \sqrt{M - E^2} + \alpha\cdot E$ and $J' = \sqrt{M' - (E')^2} + \alpha\cdot E'$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E' = E - \Delta x_i/n$ (hence $E - E' = \Delta x_i/n$) and therefore,

$$(E')^2 = E^2 - \frac{2\cdot\Delta x_i\cdot E}{n} + \frac{\Delta x_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x'_i)^2 = x_i^2 - 2\cdot\Delta x_i\cdot x_i + \Delta x_i^2$, we conclude that

$$M' = M - \frac{2\cdot\Delta x_i\cdot x_i}{n} + \frac{\Delta x_i^2}{n}.$$

Therefore, we have

$$(\sigma')^2 = M' - (E')^2 = M - \frac{2\cdot\Delta x_i\cdot x_i}{n} + \frac{\Delta x_i^2}{n} - E^2 + \frac{2\cdot\Delta x_i\cdot E}{n} - \frac{\Delta x_i^2}{n^2}.$$

Since $M - E^2 = \sigma^2$, we can rewrite this expression as follows:

$$(\sigma')^2 = \sigma^2 - \frac{2\cdot\Delta x_i\cdot x_i}{n} + \frac{\Delta x_i^2}{n} + \frac{2\cdot\Delta x_i\cdot E}{n} - \frac{\Delta x_i^2}{n^2}.$$

The inequality $J \le J'$ can be rewritten as $\sigma + \alpha\cdot E \le \sigma' + \alpha\cdot E'$. Moving $\alpha\cdot E'$ to the other side of this inequality, we conclude that

$$\sigma + \alpha\cdot(E - E') \le \sigma'.$$

Substituting the known expression for $E - E'$, we get

$$\sigma + \alpha\cdot\frac{\Delta x_i}{n} \le \sigma'.$$

Since $\Delta x_i > 0$, the left-hand side of this inequality is positive therefore, the right-hand side is also positive. Therefore, we can square both sides of this inequality and get a new inequality

$$\sigma^2 + 2\cdot\alpha\cdot\sigma\cdot\frac{\Delta x_i}{n} + \alpha^2\cdot\frac{\Delta x_i^2}{n^2} \le (\sigma')^2.$$

Substituting the above expression for $(\sigma')^2$, we get:

$$\sigma^2 + 2 \cdot \alpha \cdot \sigma \cdot \frac{\Delta x_i}{n} + \alpha^2 \cdot \frac{\Delta x_i^2}{n^2} \le \sigma^2 - \frac{2 \cdot \Delta x_i \cdot x_i}{n} + \frac{\Delta x_i^2}{n} + \frac{2 \cdot \Delta x_i \cdot E}{n} - \frac{\Delta x_i^2}{n^2}.$$

Subtracting σ^2 from both sides of the resulting inequality and dividing both sides by $\Delta x_i / n$, we conclude that

$$2 \cdot \alpha \cdot \sigma + \alpha^2 \cdot \frac{\Delta x_i}{n} \le -2 \cdot x_i + \Delta x_i + 2 \cdot E - \frac{\Delta x_i}{n.}$$

Moving all the terms containing Δx_i to the right-hand side and all other terms to the left-hand side, we conclude that

$$2 \cdot x_i - 2 \cdot E + 2 \cdot \alpha \cdot \sigma \le \Delta x_i - \frac{1+\alpha^2}{n} \cdot \Delta x_i.$$

By definition, $\Delta x_i = x_i - (E - \alpha \cdot \sigma)$, therefore, the left-hand side of this formula has the form $2 \cdot \Delta x_i$, so this formula has the form

$$2 \cdot \Delta x_i \le \Delta x_i - \frac{1+\alpha^2}{n} \cdot \Delta x_i.$$

Since $\Delta x_i > 0$, we can divide both sides of this inequality by Δx_i and conclude that $2 < 1 - (1+\alpha^2)/n$, which is impossible.

The contradiction show that our assumption that when $E - \alpha \cdot \sigma \in (\underline{x}_i, \overline{x}_i)$, the minimum can be attained for $x_i = \overline{x}_i$ is impossible. Similarly, we can prove that the minimum of the function U cannot be attained for $x_i = \underline{x}_i$. Therefore, for such i, the minimum can only be attained when $x_i = E - \alpha \cdot \sigma$.

3°. Let us now consider the case when $E - \alpha \cdot \sigma \notin (\underline{x}_i, \overline{x}_i)$. In this case, the minimum is attained either for $x_i = \overline{x}_i$ or for $x_i = \underline{x}_i$.

Let us first consider the case when the minimum is attained for $x_i = \overline{x}_i$. The fact that the minimum is attained for $x_i = \overline{x}_i$ means, in particular, that if we keep all the other values x_j the same but replace x_i by $x_i' = \underline{x}_i = x_i - 2 \cdot \Delta_i$, then the value U will not decrease. Similarly to the previous part of the proof, we will denote the values of E, U, etc., that correspond to $(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)$, by E', U', etc. In these terms, the desired inequality takes the form $U \le U'$, where $U = E + k_0 \cdot \sqrt{M - E^2}$ and $U' = E' + k_0 \cdot \sqrt{M' - (E')^2}$. It is convenient to multiply both sides of this inequality by $\alpha = 1/k_0$ and get an equivalent inequality $J \le J'$, where $J = \sqrt{M - E^2} + \alpha \cdot E$ and $J' = \sqrt{M' - (E')^2} + \alpha \cdot E'$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E' = E - 2 \cdot \Delta_i / n$ (hence $E - E' = 2 \cdot \Delta_i / n$) and therefore,

$$(E')^2 = E^2 - \frac{4 \cdot \Delta_i \cdot E}{n} + \frac{4 \cdot \Delta_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x_i')^2 = x_i^2 - 4 \cdot \Delta_i \cdot x_i + 4 \cdot \Delta x_i^2$, we conclude that

$$M' = M - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n}.$$

Therefore, we have

$$(\sigma')^2 = M' - (E')^2 = M - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n} - E^2 + \frac{4 \cdot \Delta_i \cdot E}{n} - \frac{4 \cdot \Delta_i^2}{n^2}.$$

Since $M - E^2 = \sigma^2$, we can rewrite this expression as follows:

$$(\sigma')^2 = \sigma^2 - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n} + \frac{4 \cdot \Delta_i \cdot E}{n} - \frac{4 \cdot \Delta_i^2}{n^2}.$$

The inequality $J \le J'$ can be rewritten as $\sigma + \alpha \cdot E \le \sigma' + \alpha \cdot E'$. Moving $\alpha \cdot E'$ to the other side of this inequality, we conclude that

$$\sigma + \alpha \cdot (E - E') \le \sigma'.$$

Substituting the known expression for $E - E'$, we get

$$\sigma + \alpha \cdot \frac{2 \cdot \Delta_i}{n} \le \sigma'.$$

Since $\Delta_i \ge 0$, the left-hand side of this inequality is non-negative; therefore, the right-hand side is also non-negative. Hence, we can square both sides of this inequality and get a new inequality

$$\sigma^2 + 4 \cdot \alpha \cdot \sigma \cdot \frac{\Delta_i}{n} + 4 \cdot \alpha^2 \cdot \frac{\Delta_i^2}{n^2} \le (\sigma')^2.$$

Substituting the above expression for $(\sigma')^2$, we get:

$$\sigma^2 + 4 \cdot \alpha \cdot \sigma \cdot \frac{\Delta_i}{n} + 4 \cdot \alpha^2 \cdot \frac{\Delta_i^2}{n^2} \le \sigma^2 - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n} + \frac{4 \cdot \Delta_i \cdot E}{n} - \frac{4 \cdot \Delta_i^2}{n^2}.$$

Subtracting σ^2 from both sides of the resulting inequality and dividing both sides by $4 \cdot (\Delta_i/n)$, we conclude that

$$\alpha \cdot \sigma + \alpha^2 \cdot \frac{\Delta_i}{n} \le -x_i + \Delta_i + E - \frac{\Delta_i}{n.}$$

Moving the term $\alpha \cdot \sigma$ to the right-hand side and moving all the terms from the right-hand side – except for E – to the left-hand side, we conclude that

$$x_i - \Delta_i + \frac{1 + \alpha^2}{n} \cdot \Delta_i \le E - \alpha \cdot \sigma.$$

Since $x_i = \overline{x}_i$, we thus conclude that $x_i - \Delta_i = \widetilde{x}_i$, so

$$E - \alpha \cdot \sigma \geq \widetilde{x}_i + \frac{1+\alpha^2}{n} \cdot \Delta_i$$

hence $E - \alpha \cdot \sigma \geq \widetilde{x}_i$.

We consider the case when $E - \alpha \cdot \sigma \notin (\underline{x}_i, \overline{x}_i)$, i.e., when $E - \alpha \cdot \sigma \leq \underline{x}_i$ or $E - \alpha \cdot \sigma \geq \overline{x}_i$. Since $E - \alpha \cdot \sigma > \widetilde{x}_i$, we cannot have $E - \alpha \cdot \sigma \leq \underline{x}_i$, therefore, $E - \alpha \cdot \sigma \geq \overline{x}_i$.

Similarly, when the minimum is attained for $x_i = \underline{x}_i$, we have $E - \alpha \cdot \sigma \leq \underline{x}_i$. Thus:

- when $E - \alpha \cdot \sigma \leq \underline{x}_i$, the minimum cannot be attained for $x_i = \overline{x}_i$ and therefore, it is attained when $x_i = \underline{x}_i$;
- when $\overline{x}_i \leq E - \alpha \cdot \sigma$, the minimum cannot be attained for $x_i = \underline{x}_i$ and therefore, it is attained when $x_i = \overline{x}_i$.

4°. Due to what we have proven in Parts 2° and 3° of this proof, once we know how the value $\mu \stackrel{\text{def}}{=} E - \alpha \cdot \sigma$ is located with respect to all the intervals $[\underline{x}_i, \overline{x}_i]$, we can find the optimal values of x_i:

- if $\overline{x}_i \leq \mu$, then minimum is attained when $x_i = \overline{x}_i$;
- if $\mu \leq \underline{x}_i$, then minimum is attained when $x_i = \underline{x}_i$;
- if $\underline{x}_i < \mu < \overline{x}_i$, then minimum is attained when $x_i = \mu$.

Hence, to find the minimum, we will analyze how the endpoints $\underline{x}_i$ and $\overline{x}_i$ divide the real line, and consider all the resulting sub-intervals.

Let the corresponding subinterval $[x_{(k)}, x_{(k+1)}]$ be fixed. For the i's for which $\mu \notin (\underline{x}_i, \overline{x}_i)$, the values x_i that correspond to the minimal sample variance are uniquely determined by the above formulas.

For the i's for which $\mu \in (\underline{x}_i, \overline{x}_i)$, the selected value x_i should be equal to μ. To determine this μ, we will use the fact that $\mu = E - \alpha \cdot \sigma$, where E and σ are computed by using the same value of μ.

The value E is the average of all the values x_i, i.e., the sum of all the values x_i divided by n. The sum of those values that are different from μ was denoted, in the description of the algorithm, by e_k. By using notations from the algorithm, we conclude that there are $n - n_k$ values of x_i that are equal to μ. So, the total sum of all the values x_i is equal to $e_k + (n - n_k) \cdot \mu$ and therefore, the average E is equal to

$$E = \frac{e_k + (n - n_k) \cdot \mu}{n} = \frac{e_k}{n} + \frac{n - n_k}{n} \cdot \mu.$$

Similarly, the sample second moment M is equal to:

$$M = \frac{m_k + (n - n_k) \cdot \mu^2}{n} = \frac{m_k}{n} + \frac{n - n_k}{n} \cdot \mu^2;$$

therefore,

$$\sigma^2 = M - E^2 = \left(\frac{m_k}{n} - \frac{e_k^2}{n^2}\right) - \frac{2 \cdot e_k \cdot (n - n_k)}{n^2} \cdot \mu + \left(\frac{n - n_k}{n} - \frac{(n - n_k)^2}{n^2}\right) \cdot \mu^2.$$

The coefficients at 1 and at μ^2 can be simplified, so we get

$$\sigma^2 = \frac{m_k \cdot n - e_k^2}{n^2} - \frac{2 \cdot e_k \cdot (n - n_k)}{n^2} \cdot \mu + \frac{(n - n_k) \cdot n_k}{n^2} \cdot \mu^2.$$

The condition $\mu = E - \alpha \cdot \sigma$ can be rewritten as $E - \mu = \alpha \cdot \sigma$. This equality, in its turn, is equivalent to $\mu \le E$ and $(E - \mu)^2 = \alpha^2 \cdot \sigma^2$.

The inequality $\mu \le E$ is equivalent to $\mu \cdot n_k \le e_k$. To express the second equation in terms of μ, we first take into consideration that here,

$$E - \mu = \frac{e_k}{n} + \left(\frac{n - n_k}{n} - 1\right) \cdot \mu = \frac{e_k}{n} - \frac{n_k}{n} \cdot \mu;$$

therefore,

$$(E - \mu)^2 = \frac{e_k^2}{n^2} - \frac{2 \cdot e_k \cdot n_k}{n^2} \cdot \mu + \frac{n_k^2}{n^2} \cdot \mu^2.$$

Substituting the expressions for $(E - \mu)^2$ and σ^2 into the equation $(E - \mu)^2 = \alpha^2 \cdot \sigma^2$, and multiplying both sides by n^2, we get exactly the equation given in the algorithm.

5°. To complete the proof of Theorem 2.1, we must show that this algorithm indeed requires quadratic time.

Indeed, sorting requires $O(n \cdot \log(n))$ steps, and the rest of the algorithm requires linear time ($O(n)$) for each of $2n$ subintervals, i.e., the total quadratic time.

The theorem is proven.

Proof of Theorem 3.1

Since $U = E + k_0 \cdot \sigma = k_0 \cdot J$, where $J \overset{\text{def}}{=} \sigma + \alpha \cdot E$ and $\alpha = 1/k_0$, we have $\overline{U} = k_0 \cdot \overline{J}$, where $\overline{J}$ is the upper endpoint of the interval of possible values of J. Thus, to prove that computing $\overline{U}$ is NP-hard, it is sufficient to prove that computing $\overline{J}$ is NP-hard.

To prove that the problem of computing $\overline{J}$ is NP-hard, we will prove that the known NP-hard *subset* problem $\mathcal{P}_0$ can be reduced to it. In the subset problem, given m positive integers $s_1, \ldots, s_m$, we must check whether there exist signs $\eta_i \in \{-1, +1\}$ for which the signed sum $\sum\limits_{i=1}^{m} \eta_i \cdot s_i$ equals 0.

We will show that this problem can be reduced to the problem of computing $\overline{J}$, i.e., that to every instance $(s_1, \ldots, s_m)$ of the problem $\mathcal{P}_0$, we can put into correspondence such an instance of the $\overline{J}$-computing problem that based on its solution, we can easily check whether the desired signs exist.

For that, we compute three auxiliary values

$$S \stackrel{\text{def}}{=} \frac{1}{m}\cdot\sum_{i=1}^{m} s_i^2;\quad N \stackrel{\text{def}}{=} \alpha\cdot\sqrt{\frac{2S}{1-\alpha^2}};\quad J_0 \stackrel{\text{def}}{=} (1+\alpha^2)\cdot\sqrt{\frac{S}{2\cdot(1-\alpha^2)}};$$

since $k_0 > 1$, we have $\alpha < 1$, so these definitions make sense. Then, we take $n = 2\cdot m$, $[\underline{x}_i, \overline{x}_i] = [-s_i, s_i]$ for $i = 1, 2, \ldots, m$, and $[\underline{x}_i, \overline{x}_i] = [N, N]$ for $i = m+1, \ldots, 2\cdot m$. We want to show that for the corresponding problem, we always have $\overline{J} \le J_0$, and $\overline{J} = J_0$ if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

Let us first prove that $\overline{J} \le J_0$. Since $\overline{J}$ is the upper endpoint of the interval of possible values of J, this inequality is equivalent to proving that $J \le J_0$ for all possible values J – i.e., for the values J corresponding to all possible values $x_i \in \mathbf{x}_i$.

Indeed, it is known that $V = M - E^2$, where $M \stackrel{\text{def}}{=} (1/n)\cdot\sum\limits_{i=1}^{n} x_i^2$ is the sample second moment; therefore, $J = \sqrt{M - E^2} + \alpha\cdot E$. This expression for J can be viewed as a scalar (dot) product $\vec{a}\cdot\vec{b}$ of two 2-D vectors $\vec{a} \stackrel{\text{def}}{=} (1, \alpha)$ and $\vec{b} \stackrel{\text{def}}{=} (\sqrt{M - E^2}, E)$. It is well known that for arbitrary vectors $\vec{a}$ and $\vec{b}$, we have $\vec{a}\cdot\vec{b} \le \|\vec{a}\|\cdot\|\vec{b}\|$. In our case, $\|\vec{a}\| = \sqrt{1+\alpha^2}$ and $\|\vec{b}\| = \sqrt{M}$, hence $J \le \sqrt{1+\alpha^2}\cdot\sqrt{M}$.

Since $|x_i| \le s_i$ for $i \le m$ and $x_i = N$ for $i > m$, we conclude that

$$M \le \frac{1}{2\cdot m}\cdot\sum_{i=1}^{m} x_i^2 + \frac{1}{2\cdot m}\cdot\sum_{i=m+1}^{2\cdot m} x_i^2 = \frac{1}{2}\cdot S + \frac{1}{2}\cdot N^2;$$

therefore, $J \le \sqrt{1+\alpha^2}\cdot\sqrt{(S+N^2)/2}$. Substituting the expression that defines N into this formula, we conclude that $J \le J_0$.

To complete our proof, we will show that if $J = J_0$, then $x_i = \eta_i \cdot s_i$ for $i \le m$, and $\sum\limits_{i=1}^{m} x_i = \sum\limits_{i=1}^{m} \eta_i \cdot s_i = 0$. Let us first prove that $x_i = \pm s_i$. Indeed:

- we know that $J = J_0$ and that $J_0 = \sqrt{1+\alpha^2}\cdot\sqrt{(S+N^2)/2}$, so $J = \sqrt{1+\alpha^2}\cdot\sqrt{(S+N^2)/2}$;

- we have proved that in general, $J \le \sqrt{1+\alpha^2}\cdot\sqrt{M} \le \sqrt{1+\alpha^2}\cdot\sqrt{(S+N^2)/2}$.

Therefore, $J = \sqrt{1+\alpha^2}\cdot\sqrt{(S+N^2)/2} = \sqrt{1+\alpha^2}\cdot\sqrt{M}$, hence $M = (S+N^2)/2$. If $|x_j| < s_j$ for some $j \le m$, then, from the fact that $|x_i| \le s_i$ for all $i \le m$ and $x_i = N$ for all $i > m$, we conclude that $M < (S+N^2)/2$. Thus, for every i from 1 to m, we have $|x_i| = s_i$, hence $x_i = \eta_i \cdot s_i$ for some $\eta_i \in \{-1, 1\}$.

Let us now show that $a \stackrel{\text{def}}{=} \frac{1}{m} \cdot \sum\limits_{i=1}^{m} x_i = 0$. Indeed, since $x_i = N$ for $i > m$, we have

$$E = \frac{1}{2 \cdot m} \cdot \sum_{i=1}^{m} x_i + \frac{1}{2 \cdot m} \cdot \sum_{i=m+1}^{2 \cdot m} x_i = \frac{1}{2} \cdot a + \frac{1}{2} \cdot N;$$

therefore, to prove that $a = 0$, it is sufficient to prove that $E = N/2$. The value of E can deduced from the following:

- we have just shown that in our case, $J = \sqrt{1+\alpha^2} \cdot \sqrt{M}$, where $M = (S + N^2)/2$, and
- we know that in general, $J = \vec{a} \cdot \vec{b} \le \|\vec{a}\| \cdot \|\vec{b}\| = \sqrt{1+\alpha^2} \cdot \sqrt{M}$, where the vectors $\vec{a}$ and $\vec{b}$ are defined above.

Therefore, in this case, $\vec{a} \cdot \vec{b} = \|\vec{a}\| \cdot \|\vec{b}\|$, and hence, the vectors $\vec{a} = (1, \alpha)$ and $\vec{b} = (\sqrt{M - E^2}, E)$ are parallel (proportional) to each other, i.e., $\sqrt{M - E^2}/1 = E/\alpha$ hence $E = \alpha \cdot \sqrt{M - E^2}$. From this equality, we conclude that $E > 0$ and, squaring both sides, that $E^2 = \alpha^2 \cdot (M - E^2)$ hence $(1 + \alpha^2) \cdot E^2 = \alpha^2 \cdot M = \alpha^2 \cdot (S + N^2)/2$ and $E^2 = \alpha^2 \cdot (S + N^2)/(2 \cdot (1 + \alpha^2))$. Substituting the expression that defines N into this formula, we conclude that $E^2 = N^2/4$, so, since $E > 0$, we conclude that $E = N/2$ – and therefore, that $a = 0$. The theorem is proven.

Proof of Theorem 3.2

This proof is similar to the proof of Theorem 3.1, with the only difference that we consider $J = \sigma - \alpha \cdot E$ and we take $\mathbf{x}_i = -N$ for $i > m$.

Proof of Theorems 4.1 and 4.2

We will only prove Theorem 4.1; the proof of Theorem 4.2 is practically identical.

We have already mentioned, in the proof of Theorem 2.1, that when the function $U(x_1, \ldots, x_n)$ attains its largest possible value at some point $(x_1^{\text{opt}}, \ldots, x_n^{\text{opt}})$, then, for every i, the corresponding function of one variable

$$U_i(x_i) \stackrel{\text{def}}{=} U(x_1^{\text{opt}}, \ldots, x_{i-1}^{\text{opt}}, x_i, x_{i+1}^{\text{opt}}, \ldots, x_n^{\text{opt}})$$

– the function that is obtained from $U(x_1, \ldots, x_n)$ by fixing the values of all the variables except for x_i – also attains its maximum at the value $x_i = x_i^{\text{opt}}$.

A differentiable function of one variable attains its maximum on a closed interval either at one of its endpoints or at an internal point in which its first derivative is equal to 0 and its second derivative is non-positive. We will show that for the function $U_i(x_i)$ defined on the interval $\mathbf{x}_i$, no such internal point is possible and therefore, x_i^{opt} is always equal to one of the endpoints of the interval $[\underline{x}_i, \overline{x}_i]$.

Indeed, $U = E + k_0 \cdot \sigma$. As we have mentioned in the proof of Theorem 2.1, $\sigma = \sqrt{M - E^2}$,

$$\frac{\partial E}{\partial x_i} = \frac{1}{n}; \quad \frac{\partial M}{\partial x_i} = \frac{2 \cdot x_i}{n}; \quad \frac{\partial \sigma}{\partial x_i} = \frac{x_i - E}{\sigma \cdot n}.$$

Hence, we conclude that

$$\frac{dU_i}{dx_i} = \frac{\partial U}{\partial x_i} = \frac{1}{n} + k_0 \cdot \frac{x_i - E}{\sigma \cdot n}.$$

Therefore, this first derivative is equal to 0 when $\sigma + k_0 \cdot (x_i - E) = 0$, i.e., when $x_i - E = -\alpha \cdot \sigma$ (where, as in the main text, we denoted $\alpha \stackrel{\text{def}}{=} 1/k_0$).

To get the expression for the second derivative, we differentiate the expression for the first derivative w.r.t x_i and using the above expressions for the derivatives of E and σ; as a result, we conclude that

$$\frac{d^2 U_i}{dx_i^2} = \frac{\partial^2 U}{\partial x_i^2} = \frac{k_0}{\sigma^2 \cdot n} \cdot \left(\left(1 - \frac{1}{n}\right) \cdot \sigma - (x_i - E) \cdot \frac{x_i - E}{\sigma \cdot n} \right) =$$

$$\frac{k_0}{\sigma^3 \cdot n} \cdot \left(\left(1 - \frac{1}{n}\right) \cdot \sigma^2 - \frac{1}{n} \cdot (x_i - E)^2 \right).$$

Substituting the above expression for $x_i - E = -\alpha \cdot \sigma$, we conclude that

$$\frac{d^2 U_i}{dx_i^2} = \frac{k_0}{\sigma^3 \cdot n} \cdot \left(\left(1 - \frac{1}{n}\right) - \frac{\alpha^2}{n} \right) \cdot \sigma^2.$$

Since we assumed that $1 + (1/k_0)^2 = 1 + \alpha^2 < n$, we conclude that $1 - (1/n) - (\alpha^2/n) > 0$, so the second derivative is positive and therefore, we cannot have a maximum in an internal point. The theorem is proven.

Proof of Theorems 4.3–4.6

Similarly to the case of the previous two theorems, we will prove Theorems 4.3 and 4.5; the proof of Theorems 4.4 and 4.6 is, in effect, the same.

Let us first prove that the algorithm described in Section 4 is indeed correct. Since $1 + (1/k_0)^2 < n$, we can use Theorem 4.1 and conclude that the maximum of the function U is attained when for every i, either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$. For each i, we will consider both these cases.

If the maximum is attained for $x_i = \overline{x}_i$, this means, in particular, that if we keep all the other values x_j the same but replace x_i by $x'_i = \underline{x}_i = x_i - 2 \cdot \Delta_i$, then the value U will decrease. We will denote the values of E, U, etc., that correspond to $(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)$, by E', U', etc. In these terms, the desired inequality takes the form $U \geq U'$, where $U = E + k_0 \cdot \sqrt{M - E^2}$ and $U' = E' + k_0 \cdot \sqrt{M' - (E')^2}$. Similarly to the proof of Theorem 2.1, it is convenient to multiply both sides of this inequality by $\alpha = 1/k_0$ and get an equivalent inequality $J \geq J'$, where $J = \sqrt{M - E^2} + \alpha \cdot E$ and $J' = \sqrt{M' - (E')^2} + \alpha \cdot E'$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E' = E - (2 \cdot \Delta_i)/n$ and therefore,

$$(E')^2 = E^2 - \frac{4 \cdot \Delta_i \cdot E}{n} + \frac{4 \cdot \Delta_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x_i')^2 = x_i^2 - 4 \cdot \Delta_i \cdot x_i + 4 \cdot \Delta_i^2$, we conclude that

$$M' = M - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n}.$$

Therefore, the inequality $J \geq J'$ takes the form

$$\sigma + \alpha \cdot E \geq \sqrt{M - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n} - E^2 + \frac{4 \cdot \Delta_i \cdot E}{n} - \frac{4 \cdot \Delta_i^2}{n^2}} + \alpha \cdot E'.$$

Let us simplify this expression some. First, we move the term $\alpha \cdot E'$ to the left-hand side and take into consideration that $E - E' = (2 \cdot \Delta_i)/n$; next, we take into account that $M - E^2 = \sigma^2$, so we can replace the two terms M and $-E^2$ under the square root by a single term σ^2. As a result, we arrive at the following inequality:

$$\sigma + 2 \cdot \alpha \cdot \frac{\Delta_i}{n} \geq \sqrt{\sigma^2 - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n} + \frac{4 \cdot \Delta_i \cdot E}{n} - \frac{4 \cdot \Delta_i^2}{n^2}}.$$

Since both sides of this inequality are non-negative, we can square them and get the new inequality

$$\sigma^2 + 4 \cdot \alpha \cdot \sigma \cdot \frac{\Delta_i}{n} + 4 \cdot \alpha^2 \cdot \frac{\Delta_i^2}{n^2} \geq \sigma^2 - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n} + \frac{4 \cdot \Delta_i \cdot E}{n} - \frac{4 \cdot \Delta_i^2}{n^2}.$$

If we subtract σ^2 from both sides of this inequality and divide both sides by $(4 \cdot \Delta_i)/n$, we conclude that

$$\alpha \cdot \sigma + \frac{\alpha^2}{n} \cdot \Delta_i \geq -x_i + E + \Delta_i - \frac{\Delta_i}{n}.$$

If we move terms around so that the terms proportional to x_i and Δ_i are in the left-hand side and all other terms are on the right-hand side, and take into account that since $x_i = \overline{x}_i = \widetilde{x}_i + \Delta_i$, we have $x_i - \Delta_i = \widetilde{x}_i$, we conclude that

$$\widetilde{x}_i + \Delta_i \cdot \frac{1 + \alpha^2}{n} \geq E - \alpha \cdot \sigma.$$

Similarly, if the maximum is attained for $x_i = \overline{x}_i$, this means, in particular, that if we keep all the other values x_j the same but replace x_i by $x_i' = \overline{x}_i =$

$x_i+2\cdot\Delta_i$, then the value U will decrease. We will denote the values of E, U, etc., that correspond to $(x_1,\ldots,x_{i-1},x'_i,x_{i+1},\ldots,x_n)$, by E', U', etc. In these terms, the desires inequality takes the form $U\ge U'$, where $U=E+k_0\cdot\sqrt{M-E^2}$ and $U'=E'+k_0\cdot\sqrt{M'-(E')^2}$. Similarly to the previous case, we multiply both sides of this inequality by $\alpha=1/k_0$ and get an equivalent inequality $J\ge J'$, where $J=\sqrt{M-E^2}+\alpha\cdot E$ and $J'=\sqrt{M'-(E')^2}+\alpha\cdot E'$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E'=E+(2\cdot\Delta_i)/n$ and therefore,

$$(E')^2=E^2+\frac{4\cdot\Delta_i\cdot E}{n}+\frac{4\cdot\Delta_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x'_i)^2=x_i^2+4\cdot\Delta_i\cdot x_i+4\cdot\Delta_i^2$, we conclude that

$$M'=M+\frac{4\cdot\Delta_i\cdot x_i}{n}+\frac{4\cdot\Delta_i^2}{n}.$$

Therefore, the inequality $J\ge J'$ takes the form

$$\sigma+\alpha\cdot E\ge\sqrt{M+\frac{4\cdot\Delta_i\cdot x_i}{n}+\frac{4\cdot\Delta_i^2}{n}-E^2-\frac{4\cdot\Delta_i\cdot E}{n}-\frac{4\cdot\Delta_i^2}{n^2}}+\alpha\cdot E'.$$

Let us simplify this expression. First, we move the term $\alpha\cdot E'$ to the left-hand side and take into consideration that $E-E'=-(2\cdot\Delta_i)/n$; next, we take into account that $M-E^2=\sigma^2$, so replace the two terms M and $-E^2$ under the square root by a single term σ^2. As a result, we arrive at the following inequality:

$$\sigma-2\cdot\alpha\cdot\frac{\Delta_i}{n}\ge\sqrt{\sigma^2+\frac{4\cdot\Delta_i\cdot x_i}{n}+\frac{4\cdot\Delta_i^2}{n}-\frac{4\cdot\Delta_i\cdot E}{n}-\frac{4\cdot\Delta_i^2}{n^2}}.$$

Since both sides of this inequality are non-negative, we can square them and get the new inequality

$$\sigma^2-4\cdot\alpha\cdot\sigma\cdot\frac{\Delta_i}{n}+4\cdot\alpha^2\cdot\frac{\Delta_i^2}{n^2}\ge\sigma^2+\frac{4\cdot\Delta_i\cdot x_i}{n}+\frac{4\cdot\Delta_i^2}{n}-\frac{4\cdot\Delta_i\cdot E}{n}-\frac{4\cdot\Delta_i^2}{n^2}.$$

If we subtract σ^2 from both sides of this inequality and divide both sides by $(4\cdot\Delta_i)/n$, we conclude that

$$-\alpha\cdot\sigma+\frac{\alpha^2}{n}\cdot\Delta_i\ge x_i-E+\Delta_i-\frac{\Delta_i}{n}.$$

If we move terms around so that the terms proportional to x_i and Δ_i are in the right-hand side and all other terms are on the left-hand side, and take into account that since $x_i=\underline{x}_i=\widetilde{x}_i-\Delta_i$, we have $x_i+\Delta_i=\widetilde{x}_i$, we conclude that

$$\widetilde{x}_i-\Delta_i\cdot\frac{1+\alpha^2}{n}\le E-\alpha\cdot\sigma.$$

So:

- if $x_i = \overline{x}_i$, then $E - \alpha \cdot \sigma \le \widetilde{x}_i + \Delta_i \cdot \frac{1+\alpha^2}{n}$;
- if $x_i = \underline{x}_i$, then $E - \alpha \cdot \sigma \ge \widetilde{x}_i - \Delta_i \cdot \frac{1+\alpha^2}{n}$.

Therefore, if we know the value of $E - \alpha \cdot \sigma$, then:

- if $\widetilde{x}_i + \Delta_i \cdot \frac{1+\alpha^2}{n} < E - \alpha \cdot \sigma$, then we cannot have $x_i = \overline{x}_i$ hence $x_i = \underline{x}_i$;
- similarly, if $\widetilde{x}_i - \Delta_i \cdot \frac{1+\alpha^2}{n} > E - \alpha \cdot \sigma$, then we cannot have $x_i = \underline{x}_i$ hence $x_i = \overline{x}_i$.

The only case when we do not know what value to choose is the case when

$$\widetilde{x}_i - \Delta_i \cdot \frac{1+\alpha^2}{n} \le E - \alpha \cdot \sigma \le \widetilde{x}_i + \Delta_i \cdot \frac{1+\alpha^2}{n},$$

i.e., when the value $E - \alpha \cdot \sigma$ belongs to the i-th narrowed interval; in this case, we can, in principle, have both $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$. Thus, the algorithm is indeed correct.

Let us prove that this algorithm requires quadratic time. Indeed, once we know where E is with respect to the endpoints of all narrowed intervals, we can determine the values of all optimal x_i – except for those that are within this narrowed interval. Since we consider the case when no more than C narrowed intervals can have a common point, we have no more than C undecided values x_i. Trying all possible combinations of lower and upper endpoints for these $\le C$ values requires $\le 2^C$ steps. For each zone and for each of these combinations, we need a linear time ($O(n)$) to compute U. Thus, for each zone, we need $O(2^C \cdot n)$ computational steps. There are $O(n)$ zones, so the overall number of steps is $O(2^C \cdot n^2)$. Since C is a constant, the overall number of steps is thus $O(n^2)$.

The theorem is proven.

7 Proof of Theorem 5.1

1°. Let us first show that the algorithm described is indeed correct.

Let us first consider the case when all the intervals intersect. We know that the variance $V = M - E^2$ is always non-negative; therefore, $M \ge E^2$ and $R \ge 1$; hence $\underline{R} \ge 1$. If all the intervals have a common point, that it is possible that all the values x_i are equal to this common point; in this case, $V = 0$ hence $R = 1$. Thus, in this case, $\underline{R} = 1$.

Let us now consider the case when the intersection of n intervals is empty. A differentiable function of one variable attains its minimum on a closed interval either at one of its endpoints or at an internal point in which its first derivative is equal to 0. It is known that $M \stackrel{\text{def}}{=} (1/n) \cdot \sum\limits_{i=1}^{n} x_i^2$ and $E \stackrel{\text{def}}{=} (1/n) \cdot \sum\limits_{i=1}^{n} x_i$. Here,

$$\frac{\partial E}{\partial x_i} = \frac{1}{n}; \quad \frac{\partial M}{\partial x_i} = \frac{2 \cdot x_i}{n};$$

therefore,

$$\frac{\partial}{\partial x_i}\left(\frac{M}{E^2}\right) = \frac{\dfrac{\partial M}{\partial x_i} \cdot E^2 - M \cdot 2 \cdot E \cdot \dfrac{\partial E}{\partial x_i}}{E^4} = \frac{1}{E^4} \cdot \left(\frac{2 \cdot x_i}{n} \cdot E^2 - M \cdot 2 \cdot E \cdot \frac{1}{n}\right) =$$

$$\frac{2}{n \cdot E^3} \cdot (E \cdot x_i - M).$$

Therefore, this derivative is equal to 0 when $E \cdot x_i - M = 0$, i.e., when $x_i = M/E$.

Thus, for the optimal values $x_1, \ldots, x_n$ for which M/E^2 attains its minimum, for every i, we have either $x_i = \underline{x}_i$, or $x_i = \overline{x}_i$, or $x_i = M/E$.

2°. Let us show that if the open interval $(\underline{x}_i, \overline{x}_i)$ contains the value M/E, then the minimum of the function cannot be attained at points $\overline{x}_i$ or $\underline{x}_i$ and therefore, has to be attained at the value M/E.

Let us show that the minimum cannot be attained for $x_i = \underline{x}_i$ (for $x_i = \overline{x}_i$, the proof is similar). We will prove this impossibility by reduction to a contradiction: namely, we assume that the minimum is attained for $x_i = \underline{x}_i$, and we will deduce a contradiction from this assumption.

The fact that the minimum is attained for $x_i = \underline{x}_i$ means, in particular, that if we keep all the other values the same but replace x_i by $x'_i = M/E$, then the value of M/E^2 cannot decrease. Let us denote the change in x_i by $\Delta x_i \stackrel{\text{def}}{=} x'_i - x_i$; clearly, $\Delta x_i > 0$. We will denote the values of E, M, etc., that correspond to $(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)$, by E', M', etc. In these terms, the fact that the minimum is attained for $x_i = \underline{x}_i$ means that $M/E^2 \le M'/(E')^2$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E' = E + \Delta x_i / n$ and therefore,

$$(E')^2 = E^2 + \frac{2 \cdot \Delta x_i \cdot E}{n} + \frac{\Delta x_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x'_i)^2 = x_i^2 + 2 \cdot \Delta x_i \cdot x_i + \Delta x_i^2$, we conclude that

$$M' = M + \frac{2 \cdot \Delta x_i \cdot x_i}{n} + \frac{\Delta x_i^2}{n}.$$

We have concluded that $M/E^2 \le M'/(E')^2$, i.e., that $M \cdot (E')^2 < M' \cdot E^2$. Substituting the above expression for $(E')^2$ and M' into this inequality, we get

$$M \cdot E^2 + M \cdot \frac{2 \cdot E \cdot \Delta x_i}{n} + M \cdot \frac{\Delta x_i^2}{n^2} \le M \cdot E^2 + \frac{2 \cdot \Delta x_i \cdot x_i}{n} \cdot E^2 + \frac{\Delta x_i^2}{n} \cdot E^2.$$

Subtracting $M \cdot E^2$ from both sides, we conclude that

$$M \cdot \frac{2 \cdot E \cdot \Delta x_i}{n} + M \cdot \frac{\Delta x_i^2}{n^2} \le \frac{2 \cdot \Delta x_i \cdot x_i}{n} \cdot E^2 + \frac{\Delta x_i^2}{n} \cdot E^2.$$

Dividing both sides by $\Delta x_i/n$, we conclude that

$$2 \cdot M \cdot E + M \cdot \frac{\Delta x_i}{n} \le 2 \cdot x_i \cdot E^2 + \Delta x_i \cdot E^2.$$

Substituting $\Delta x_i = (M/E) - x_i$, we get

$$2 \cdot M \cdot E + \frac{M \cdot \left(\dfrac{M}{E} - x_i\right)}{n} \le 2 \cdot x_i \cdot E^2 + \frac{M}{E} \cdot E^2 - x_i \cdot E^2.$$

Here, $(M/E) \cdot E^2 = M \cdot E$; subtracting $M \cdot E$ from both sides, we conclude that

$$M \cdot E + \frac{M}{n} \cdot \left(\frac{M}{E} - x_i\right) \le x_i \cdot E^2.$$

Moving the term $M \cdot E$ to the right-hand side, we conclude that

$$\frac{M}{n} \cdot \left(\frac{M}{E} - x_i\right) \le \left(x_i - \frac{M}{E}\right) \cdot E^2.$$

However, since $x_i = \underline{x}_i$ and $M/E \in (\underline{x}_i, \overline{x}_i)$, we conclude that $\dfrac{M}{n} \cdot \left(\dfrac{M}{E} - x_i\right)$ is positive and $\left(x_i - \dfrac{M}{E}\right) \cdot E^2$ is negative, so the above inequality is impossible.

This contradiction shows that when $M/E \in (\underline{x}_i, \overline{x}_i)$, the minimum of the function M/E^2 cannot be attained for $x_i = \underline{x}_i$. Similarly, we can prove that in this case, the minimum of the function M/E^2 cannot be attained for $x_i = \overline{x}_i$. Therefore, for such i, the minimum can only be attained when $x_i = M/E$.

3°. Let us now consider the case when $\dfrac{M}{E} \not\in (\underline{x}_i, \overline{x}_i)$. In this case, the minimum is attained either for $x_i = \overline{x}_i$ or for $x_i = \underline{x}_i$. Let us first consider the case when the minimum is attained for $x_i = \underline{x}_i$. The fact that the minimum is attained for $x_i = \underline{x}_i$ means, in particular, that if we keep all the other values the same but replace x_i by $x_i' = \overline{x}_i = x_i + 2 \cdot \Delta_i$, then the value M/E^2 will not decrease. Similarly to the previous part of the proof, we will denote the values of E, M, etc., that correspond to $(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)$, by E', M', etc. In these terms, the fact that the minimum is attained for $x_i = \underline{x}_i$ means that $M/E^2 \le M'/(E')^2$, i.e., that $M \cdot (E')^2 \le M' \cdot E^2$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E' = E + 2 \cdot \Delta_i/n$ and therefore,

$$(E')^2 = E^2 + \frac{4 \cdot \Delta_i \cdot E}{n} + \frac{4 \cdot \Delta_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x_i')^2 = x_i^2 + 4 \cdot \Delta_i \cdot x_i + 4 \cdot \Delta_i^2$, we conclude that

$$M' = M + \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n}.$$

Substituting the expressions for $(E')^2$ and M into the inequality $M \cdot (E')^2 \le M' \cdot E^2$, we get

$$M \cdot E^2 + M \cdot \frac{4 \cdot \Delta_i \cdot E}{n} + M \cdot \frac{4 \cdot \Delta_i^2}{n^2} \le M \cdot E^2 + \frac{4 \cdot \Delta_i \cdot x_i}{n} \cdot E^2 + \frac{4 \cdot \Delta_i^2}{n} \cdot E^2.$$

Subtracting $M \cdot E^2$ from both sides and dividing the resulting expressions by $\frac{4 \cdot \Delta_i}{n \cdot E^2}$, we get

$$\frac{M}{E} + \frac{M}{E^2} \cdot \frac{\Delta_i}{n} \le x_i + \Delta_i.$$

Since $x_i = \underline{x}_i$, we thus conclude that $x_i + \Delta_i = \widetilde{x}_i$, so

$$\widetilde{x}_i \ge \frac{M}{E} \cdot \left(1 + \frac{\Delta_i}{E \cdot n}\right).$$

Hence, if $\widetilde{x}_i < \frac{M}{E} \cdot \left(1 + \frac{\Delta_i}{E \cdot n}\right)$ then $x_i = \overline{x}_i$.

4°. Let us consider the case when the minimum is attained for $x_i = \overline{x}_i$. The fact that the minimum is attained for $x_i = \overline{x}_i$ means, in particular, that if we keep all the other values the same but replace x_i by $x_i' = \underline{x}_i = x_i - 2 \cdot \Delta_i$, then the value M/E^2 will not decrease. Similarly to the previous part of the proof, we will denote the values of E, M, etc., that correspond to $(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)$, by E', M', etc. In these terms, the fact that the minimum is attained for $x_i = \underline{x}_i$ means that $M/E^2 \le M'/(E')^2$, i.e., that $M \cdot (E')^2 \le M' \cdot E^2$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E' = E - 2 \cdot \Delta_i/n$ (hence $E - E' = 2 \cdot \Delta_i/n$) and therefore,

$$(E')^2 = E^2 - \frac{4 \cdot \Delta_i \cdot E}{n} + \frac{4 \cdot \Delta_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x_i')^2 = x_i^2 - 4 \cdot \Delta_i \cdot x_i + 4 \cdot \Delta_i^2$, we conclude that

$$M' = M - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n}.$$

Substituting the expressions for $(E')^2$ and M into the inequality $M \cdot (E')^2 \le M' \cdot E^2$, we get

$$M \cdot E^2 - M \cdot \frac{4 \cdot \Delta_i \cdot E}{n} + M \cdot \frac{4 \cdot \Delta_i^2}{n^2} \le M \cdot E^2 - \frac{4 \cdot \Delta_i \cdot x_i}{n} \cdot E^2 + \frac{4 \cdot \Delta_i^2}{n} \cdot E^2.$$

Subtracting $M \cdot E^2$ from both sides and dividing the resulting expressions by $\frac{4 \cdot \Delta_i}{n \cdot E^2}$, we get

$$-\frac{M}{E} + \frac{M}{E^2} \cdot \frac{\Delta_i}{n} \le -x_i + \Delta_i.$$

Since $x_i = \overline{x}_i$, we thus conclude that $x_i - \Delta_i = \widetilde{x}_i$, so

$$-\widetilde{x}_i \ge -\frac{M}{E} \cdot \left(1 - \frac{\Delta_i}{E \cdot n}\right).$$

Therefore, we have

$$\widetilde{x}_i \le \frac{M}{E} \cdot \left(1 - \frac{\Delta_i}{E \cdot n}\right).$$

Hence, if $\widetilde{x}_i > \frac{M}{E} \cdot \left(1 - \frac{\Delta_i}{E \cdot n}\right)$ then $x_i = \underline{x}_i$.

5°. Thus:

- when $\widetilde{x}_i = \frac{M}{E}$, we have $\frac{M}{E} \in [\underline{x}_i, \overline{x}_i]$ and therefore the minimum is attained at $x_i = \frac{M}{E}$,
- when $\widetilde{x}_i > \frac{M}{E}$, we have $\widetilde{x}_i > \frac{M}{E} \cdot \left(1 - \frac{\Delta_i}{E \cdot n}\right)$ and therefore the minimum is attained at $x_i = \underline{x}_i$,
- when $\widetilde{x}_i < \frac{M}{E}$, we have $\widetilde{x}_i < \frac{M}{E} \cdot \left(1 + \frac{\Delta_i}{E \cdot n}\right)$ and therefore the minimum is attained at $x_i = \overline{x}_i$.

Hence, to find the values x_i that minimize M/E^2, we must find out where $\lambda \overset{\text{def}}{=} M/E$ is in comparison with each of the intervals $[\underline{x}_i, \overline{x}_i]$. In other words, it is sufficient to separately consider all the zones into which the values $\underline{x}_i$ and $\overline{x}_i$ divide the real line.

Let a zone $[x_{(k)}, x_{(k+1)}]$ be fixed. For the i's for which $\lambda \notin (\underline{x}_i, \overline{x}_i)$, the values x_i that correspond to the minimal value of M/E^2 are uniquely determined by the above formulas.

For the i's for which $\lambda \in (\underline{x}_i, \overline{x}_i)$, the selected value x_i should be equal to λ. To determine this λ, we will use the fact that $\lambda = M/E$, where E and M are computed by using the same value of λ.

The value E is the average of all the values x_i, i.e., the sum of all the values x_i divided by n. The sum of those values that are different from λ was denoted, in the description of the algorithm, by e_k. By using notations from the algorithm, we conclude that there are $n - n_k$ values of x_i that are equal to λ. So, the total sum of all the values x_i is equal to $e_k + (n - n_k) \cdot \lambda$ and therefore, the average E is equal to

$$E = \frac{e_k + (n - n_k) \cdot \lambda}{n} = \frac{e_k}{n} + \frac{n - n_k}{n} \cdot \lambda.$$

Similarly, the sample second moment M is equal to:

$$M = \frac{m_k + (n - n_k) \cdot \lambda^2}{n} = \frac{m_k}{n} + \frac{n - n_k}{n} \cdot \lambda^2;$$

therefore,

$$\lambda = \frac{M}{E} = \frac{(m_k + (n - n_k) \cdot \lambda^2)/n}{(e_k + (n - n_k) \cdot \lambda)/n} = \frac{m_k + (n - n_k) \cdot \lambda^2}{e_k + (n - n_k) \cdot \lambda}.$$

Multiplying both sides of the equation by $e_k + (n - n_k) \cdot \lambda$, we get

$$\lambda \cdot (e_k + (n - n_k) \cdot \lambda) = m_k + (n - n_k) \cdot \lambda^2,$$

i.e.,

$$\lambda \cdot e_k + (n - n_k) \cdot \lambda^2 = m_k + (n - n_k) \cdot \lambda^2,$$

hence

$$\lambda = \frac{m_k}{e_k}.$$

Substituting this expression for λ into the expressions $E = (e_k + (n - n_k) \cdot \lambda)/n$ and $M = (m_k + (n - n_k) \cdot \lambda^2)/n$, we can thus find the minimum of the function M/E^2.

The theorem is proven.

8 Proof of Theorem 5.2

This theorem was, in effect, proven when we proved Theorem 5.1: indeed, similar to the minimum, the maximum can also be attained either at one of the endpoints, or at the point where the partial derivative is equal to 0. However, we have already shown that in this point, $x_i = M/E$, and that at this point, the function $R = M/E^2$ attains its minimum.

9 Proof of Theorem 5.3

1°. If the maximum is attained for $x_i = \underline{x}_i$, this means, in particular, that if we keep all the other values the same but replace x_i by $x'_i = \overline{x}_i = x_i + 2 \cdot \Delta_i$, then the value R will decrease. We will denote the values of E, M, etc., that correspond to $(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)$, by E', M', etc. In these terms, the desires inequality takes the form $R \geq R'$, where $R = M/E^2$ and $R' = M'/(E')^2$, hence $M \cdot (E')^2 \geq M' \cdot E^2$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E' = E + (2 \cdot \Delta_i)/n$ and therefore,

$$(E')^2 = E^2 + \frac{4 \cdot \Delta_i \cdot E}{n} + \frac{4 \cdot \Delta_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x_i')^2 = x_i^2 + 4 \cdot \Delta_i \cdot x_i + 4 \cdot \Delta_i^2$, we conclude that

$$M' = M + \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n}.$$

Substituting the values for $(E')^2$ and M into the inequality $M \cdot (E')^2 \geq M' \cdot E^2$, we get

$$M \cdot E^2 + M \cdot \frac{4 \cdot \Delta_i \cdot E}{n} + M \cdot \frac{4 \cdot \Delta_i^2}{n^2} \geq M \cdot E^2 + \frac{4 \cdot \Delta_i \cdot x_i}{n} \cdot E^2 + \frac{4 \cdot \Delta_i^2}{n} \cdot E^2.$$

Subtracting $M \cdot E^2$ from both sides and dividing both sides by $\dfrac{4 \cdot \Delta_i}{n \cdot E^2}$, we get

$$\frac{M}{E} + \frac{M}{E^2} \cdot \frac{\Delta_i}{n} \geq x_i + \Delta_i.$$

Since $x_i = \underline{x}_i$, we thus conclude that $x_i + \Delta_i = \widetilde{x}_i$, so

$$\widetilde{x}_i \leq \frac{M}{E} \cdot \left(1 + \frac{\Delta_i}{E \cdot n}\right).$$

We have already shown, in the proof of Theorem 5.2, that the maximum of the function M/E^2 can only be attained at the endpoints. Hence, if $\widetilde{x}_i > \dfrac{M}{E} \cdot \left(1 + \dfrac{\Delta_i}{E \cdot n}\right)$ then $x_i = \overline{x}_i$.

2°. Similarly, if the maximum is attained for $x_i = \overline{x}_i$, this means, in particular, that if we keep all the other values the same but replace x_i by $x_i' = \underline{x}_i = x_i - 2 \cdot \Delta_i$, then the value R will decrease. We will denote the values of E, M, etc., that correspond to $(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)$, by E', M', etc. In these terms, the desired inequality takes the form $R \geq R'$, where $R = M/E^2$ and $R' = M'/(E')^2$, i.e., $M \cdot (E')^2 \geq M' \cdot E^2$.

By definition of E as the arithmetic average of the values x_i, we conclude that $E' = E - (2 \cdot \Delta_i)/n$ and therefore,

$$(E')^2 = E^2 - \frac{4 \cdot \Delta_i \cdot E}{n} + \frac{4 \cdot \Delta_i^2}{n^2}.$$

Similarly, by definition, the sample second moment M is the average of the squares x_i^2; since $(x_i')^2 = x_i^2 - 4 \cdot \Delta_i \cdot x_i + 4 \cdot \Delta_i^2$, we conclude that

$$M' = M - \frac{4 \cdot \Delta_i \cdot x_i}{n} + \frac{4 \cdot \Delta_i^2}{n}.$$

Substituting the above expressions for $(E')^2$ and M into the inequality $M \cdot (E')^2 \geq M' \cdot E^2$, we get

$$M \cdot E^2 - M \cdot \frac{4 \cdot \Delta_i \cdot E}{n} + M \cdot \frac{4 \cdot \Delta_i^2}{n^2} \geq M \cdot E^2 - \frac{4 \cdot \Delta_i \cdot x_i}{n} \cdot E^2 + \frac{4 \cdot \Delta_i^2}{n} \cdot E^2.$$

Subtracting $M \cdot E^2$ from both sides and dividing both sides by $\dfrac{4 \cdot \Delta_i}{n \cdot E^2}$, we get

$$-\frac{M}{E} + \frac{M}{E^2} \cdot \frac{\Delta_i}{n} \geq -x_i + \Delta_i.$$

Since $x_i = \overline{x}_i$, we thus conclude that $x_i - \Delta_i = \widetilde{x}_i$, so

$$\widetilde{x}_i \geq \frac{M}{E} \cdot \left(1 - \frac{\Delta_i}{E \cdot n}\right).$$

Hence, if $\widetilde{x}_i < \dfrac{M}{E} \cdot \left(1 - \dfrac{\Delta_i}{E \cdot n}\right)$ then $x_i = \underline{x}_i$.

3°. So:

- if $x_i = \underline{x}_i$, then $\widetilde{x}_i \leq \dfrac{M}{E} \cdot \left(1 + \dfrac{\Delta_i}{E \cdot n}\right)$;
- if $x_i = \overline{x}_i$, then $\widetilde{x}_i \geq \dfrac{M}{E} \cdot \left(1 - \dfrac{\Delta_i}{E \cdot n}\right)$.

Therefore, if we know the value of M/E, then:

- if $\dfrac{\widetilde{x}_i}{1 + \frac{\Delta_i}{E \cdot n}} > \dfrac{M}{E}$, then we cannot have $x_i = \underline{x}_i$ hence $x_i = \overline{x}_i$;
- if $\dfrac{\widetilde{x}_i}{1 - \frac{\Delta_i}{E \cdot n}} < \dfrac{M}{E}$, then we cannot have $x_i = \overline{x}_i$ hence $x_i = \underline{x}_i$;

The only case when we do not know what value to choose is the case when

$$\frac{\widetilde{x}_i}{1 + \frac{\Delta_i}{\underline{E} \cdot n}} \leq M/E \leq \frac{\widetilde{x}_i}{1 - \frac{\Delta_i}{\underline{E} \cdot n}},$$

i.e., when the value M/E belongs to the i-th narrowed interval; in this case, we can, in principle, have both $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$. Thus, the algorithm is indeed correct.

Let us prove that this algorithm requires quadratic time. Indeed, once we know where E is with respect to the endpoints of all narrowed intervals, we can determine the values of all optimal x_i – except for those that are within this narrowed interval. Since we consider the case when no more than C narrowed intervals can have a common point, we have no more than C undecided values x_i. Trying all possible combinations of lower and upper endpoints for these $\leq C$ values requires $\leq 2^C$ steps. For each zone and for each of these combinations, we need a linear time ($O(n)$) to compute R. Thus, for each zone, we need $O(2^C \cdot n)$ computational steps. There are $O(n)$ zones, so the overall number of steps is $O(2^C \cdot n^2)$. Since C is a constant, the overall number of steps is thus $O(n^2)$.

The theorem is proven.

Conclusions

In many application areas, it is important to detect outliers. Traditional engineering approach to outlier detection is that we start with some "normal" values $x_1, \ldots, x_n$, compute the sample average E, the sample standard variation σ, and then mark a value x as an outlier if x is outside the k_0-sigma interval $[E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$ (for some pre-selected parameter k_0).

In real life, we often have only interval ranges $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ for the normal values $x_1, \ldots, x_n$. For different values $x_i \in \mathbf{x}_i$, we get different values of $L \overset{\text{def}}{=} E - k_0 \cdot \sigma$ and $U \overset{\text{def}}{=} E + k_0 \cdot \sigma$ – and thus, different k_0-sigma intervals $[L, U]$. We can therefore identify *guaranteed* outliers as values that are outside *all* k_0-sigma intervals, and *possible* outliers as values that are outside *some* k_0-sigma intervals. To detect guaranteed and possible outliers, we must therefore be able to compute the *range* $\mathbf{L} = [\underline{L}, \overline{L}]$ of possible values of L and the range $\mathbf{U} = [\underline{U}, \overline{U}]$ of possible values of U.

In our previous papers [3, 4], we have shown how to compute the intervals $\mathbf{E} = [\underline{E}, \overline{E}]$ and $[\underline{\sigma}, \overline{\sigma}]$ of possible values for E and σ. In principle, we can combine these intervals and conclude, e.g., that L always belongs to the interval $\mathbf{E} - k_0 \cdot [\underline{\sigma}, \overline{\sigma}]$. However, the resulting interval for L is *wider* than the actual range – wider because the values E and σ are computed based on the same inputs $x_1, \ldots, x_n$ and are, therefore, not independent from each other.

If, instead of the actual ranges for L and U, we use wider intervals, we may miss some outliers. It is therefore important to compute the *exact* ranges for L and U.

In this paper, we showed that computing these ranges is, in general, NP-hard, and we provided efficient algorithms that compute these ranges under reasonable conditions.

Once a value is identified as an outlier for a fixed k_0, we also show how to find out to what degree this value is an outlier, i.e., what is the largest value k_0 for which this value is an outlier.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209 and grant NCC2-1232, by the Future Aerospace Science and Technology

Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant numbers F49620-00-1-0365, by NSF grants CDA-9522207, EAR-0112968, EAR-0225670, and 9710940 Mexico/Conacyt and IEEE/ACM SC2001 and SC2002 Minority Serving Institutions Participation Grants.

This work was supported (in part) by Small Business Innovation Research grant 9R44CA81741 to Applied Biomathematics from the National Cancer Institute (NCI), a component of the National Institutes of Health (NIH). The opinions expressed herein are those of the authors and not necessarily those of the NCI or the NIH.

This work was partly supported by a research grant from Sandia National Laboratories as part of the Department of Energy Accelerated Strategic Computing Initiative (ASCI).

References

[1] J. Devore and R. Peck, *Statistics: the Exploration and Analysis of Data*, Duxbury, Pacific Grove, California, 1999.

[2] C. Ferregut, R. A. Osegueda, and A. Nuñez (eds.), *Proceedings of the International Workshop on Intelligent NDE Sciences for Aging and Futuristic Aircraft*, El Paso, TX, September 30–October 2, 1997.

[3] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpre, and M. Aviles, "Computing Variance for Interval Data is NP-Hard", *ACM SIGACT News*, 2002, Vol. 33, No. 2, pp. 108–118.

[4] S. Ferson, L. Ginzburg, V. Kreinovich, and M. Aviles, "Exact Bounds on Sample Variance of Interval Data", *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, Toronto, Canada, May 23–25, 2002, pp. 67–69.

[5] M. Goodchild and S. Gopal, Accuracy of Spatial Databases, Taylor & Francis, London, 1989.

[6] X. E. Gros, *NDT Data Fusion*, J. Wiley, London, 1997.

[7] O. Kosheleva, S. Cabrera, R. Osegueda, S. Nazarian, D. L. George, M. J. George, V. Kreinovich, and K. Worden, "Case study of non-linear inverse problems: mammography and non-destructive evaluation", In: A. Mohamad-Djafari (ed.), *Bayesian Inference for Inverse Problems, Proceedings of the SPIE/International Society for Optical Engineering*, Vol. 3459, San Diego, CA, 1998, pp. 128–135.

[8] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational complexity and feasibility of data processing and interval computations*, Kluwer, Dordrecht, 1997.

[9] V. Kreinovich, L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg, "Outlier Detection Under Interval Uncertainty: Algorithmic Solvability and Computational Complexity", *Proceedings of the 4-th International Conference on Large-Scale Scientific Computations*, Sozopol, Bulgaria, June 4–8, 2003, Springer Lecture Notes in Computer Science (to appear).

[10] V. Kreinovich, P. Patangay, L. Longpré, S. A. Starks, C. Campos, S. Ferson, and L. Ginzburg, "Outlier Detection Under Interval and Fuzzy Uncertainty: Algorithmic Solvability and Computational Complexity", *Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society NAFIPS'2003*, Chicago, Illinois, July 24–26, 2003 (to appear).

[11] M. McCain and C. William, "Integrating Quality Assurance into the GIS Project Life Cycle", *Proceedings of the 1998 ESRI Users Conference* http://www.dogcreek.com/html/documents.html

[12] R. Osegueda, V. Kreinovich, L. Potluri, and R. Aló, "Non-Destructive Testing of Aerospace Structures: Granularity and Data Mining Approach", *Proceedings of FUZZ-IEEE'2002*, Honolulu, Hawaii, May 12–17, 2002, Vol. 1, pp. 685–689.

[13] R. A. Osegueda, S. R. Seelam, A. C. Holguin, V. Kreinovich, and C.-W. Tao, "Statistical and Dempster-Shafer Techniques in Testing Structural Integrity of Aerospace Structures", *International Journal of Uncertainty, Fuzziness, Knowledge-Based Systems (IJUFKS)*, 2001, Vol. 9, No. 6, pp. 749–758.

[14] S. Rabinovich, Measurement Errors: Theory and Practice, American Institute of Physics, New York, 1993.

[15] L. Scott, "Identification of GIS Attribute Error Using Exploratory Data Analysis", *Professional Geographer*, 1994, Vol. 46, No. 3, pp. 378–386.

[16] S. A. Vavasis, *Nonlinear optimization: complexity issues*, Oxford University Press, N.Y., 1991.

[17] H. M. Wadsworth Jr. (editor), *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., N.Y., 1990.

[18] Q. Wen, A. Q. Gates, J. Beck, V. Kreinovich, and G. R. Keller, "Towards automatic detection of erroneous measurement results in a gravity database", *Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference*, Tucson, Arizona, October 7–10, 2001, pp. 2170–2175.

[19] K. Worden, R. Osegueda, C. Ferregut, S. Nazarian, D. L. George, M. J. George, V. Kreinovich, O. Kosheleva, and S. Cabrera, "Interval Methods in Non-Destructive Testing of Aerospace Structures and in Mammography", *International Conference on Interval Methods and their Application in Global Optimization (INTERVAL'98), April 20–23, Nanjing, China, Abstracts*, pp. 152–154, 1998.

[20] K. Worden, R. Osegueda, C. Ferregut, S. Nazarian, E. Rodriguez, D. L. George, M. J. George, V. Kreinovich, O. Kosheleva, and S. Cabrera, "Interval Approach to Non-Destructive Testing of Aerospace Structures and to Mammography", In: G. Alefeld and R. A. Trejo (eds.), *Interval Computations and its Applications to Reasoning Under Uncertainty, Knowledge Representation, and Control Theory. Proceedings of MEXICON'98, Workshop on Interval Computations, 4th World Congress on Expert Systems*, México City, México, 1998.

[21] K. Worden, R. Osegueda, C. Ferregut, S. Nazarian, D. L. George, M. J. George, V. Kreinovich, O. Kosheleva, and S. Cabrera, "Interval Methods in Non-Destructive Testing of Material Structures", *Reliable Computing*, 2001, Vol. 7, No. 4, pp. 341–352.