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. The 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., [8, 20, 21]); in geophysics, abnormal values may indicate a mineral deposit or an erroneous measurement result (see, e.g., [6, 12, 16, 19]); in structural integrity testing, abnormal values may indicate faults in a structure (see, e.g., [3, 7, 8, 13, 14, 20, 21, 22]), etc.

The traditional engineering approach to outlier detection (see, e.g., [2, 15, 18]) is as follows:

- first, we collect measurement results $x_1, \ldots, x_n$ corresponding to normal situations;
- then, we compute the sample average $E \stackrel{\text{def}}{=} \frac{x_1 + \ldots + x_n}{n}$ of these normal values and the (sample) standard deviation $\sigma = \sqrt{V}$, where $V \stackrel{\text{def}}{=} \frac{(x_1 - E)^2 + \ldots + (x_n - E)^2}{n}$;
- 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 [10, 11].

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 [4, 5], 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 U 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.

As an example that we may lose precision by combining intervals for E and σ, let us consider the case when $\mathbf{x}_1 = \mathbf{x}_2 = [0, 1]$ and $k_0 = 2$. In this case, the range $\mathbf{E}$ of $E = (x_1 + x_2)/2$ is equal to $[0, 1]$, where the largest value 1 is attained only if $x_1 = x_2 = 1$. For the variance, we have $V = ((x_1 - E)^2 + (x_2 - E)^2)/2 = (x_1 - x_2)^2/4$; so, the range $\mathbf{V}$ of V is $[0, 0.25]$ and, correspondingly, the range for $\sigma = \sqrt{V}$ is $[0, 0.5]$. The largest value $\sigma = 0.5$ is only attained in two cases: when $x_1 = 0$ and $x_2 = 1$, and when $x_1 = 1$ and $x_2 = 0$. When we simply combine the intervals, we conclude that $U \in [0, 1] + 2 \cdot [0, 0.5] = [0, 2]$. However, it is easy to see that U cannot be equal to 2:

- The only way for U to be equal to 2 is when both E and σ attain their largest values: $E = 1$ and $\sigma = 0.5$.
- However, the only pair on which the mean E attains its largest value 1 is $x_1 = x_2 = 1$, and for this pair, $\sigma = 0$.

So, the actual range of U must be narrower than the result $[0, 2]$ of combining intervals for E and σ.

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 algorithms $\underline{\mathcal{A}}_U$ for computing $\underline{U}$ and $\overline{\mathcal{A}}_L$ for computing $\overline{L}$ are as follows:

- In both algorithms, first, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(2n)}$; 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

$$e_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \geq x_{(k+1)}} \underline{x}_i + \sum_{j:\overline{x}_j \leq x_{(k)}} \overline{x}_j,$$

$$m_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \geq x_{(k+1)}} (\underline{x}_i)^2 + \sum_{j:\overline{x}_j \leq x_{(k)}} (\overline{x}_j)^2,$$

and $n_k =$ the total number of such i's and j's. Then, we solve the quadratic equation

$$A_k - B_k \cdot \mu + C_k \cdot \mu^2 = 0,$$

where

$$A_k \stackrel{\text{def}}{=} e_k^2 \cdot (1+\alpha^2) - \alpha^2 \cdot m_k \cdot n; \quad \alpha \stackrel{\text{def}}{=} 1/k_0,$$

$$B_k \stackrel{\text{def}}{=} 2 \cdot e_k \cdot \left((1+\alpha^2) \cdot n_k - \alpha^2 \cdot n\right); \quad C_k \stackrel{\text{def}}{=} n_k \cdot \left((1+\alpha^2) \cdot n_k - \alpha^2 \cdot n\right).$$

For computing $\underline{U}$, we select only those solutions for which $\mu \cdot n_k \leq e_k$ and $\mu \in [x_{(k)}, x_{(k+1)}]$; for computing $\overline{L}$, we select only those solutions for which $\mu \cdot n_k \geq e_k$ and $\mu \in [x_{(k)}, x_{(k+1)}]$. For each selected solution, we compute the values of

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

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

- Finally, if we are computing $\underline{U}$, we return the smallest of the values U_k; if we are computing $\overline{L}$, we return the smallest of the values L_k.

Theorem 2.1. *The algorithms $\underline{\mathcal{A}}_U$ and $\overline{\mathcal{A}}_L$ always compute $\underline{U}$ and $\overline{L}$ in quadratic time.*

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

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.*

Comments.

- For interval data, the NP-hardness of computing the upper bound for σ was proven in [4, 5]. A general overview of NP-hardness of computational problems in interval context is given in [9].
- The proofs of Theorems 4.1 and 4.2 show that the decision problems related to the computation of $\underline{L}$ and $\overline{U}$ are NP-complete. Therefore, NP-hardness of the computational problems does not mean that the problems are located somewhere higher in the polynomial hierarchy.

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 and the minimum of the function L on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$ are attained at 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 $\overline{U}$ and $\underline{L}$) 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") $\widetilde{x}_i \stackrel{\text{def}}{=} (\underline{x}_i + \overline{x}_i)/2$ 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 $\Delta_i \stackrel{\text{def}}{=} (\underline{x}_i - \overline{x}_i)/2$ is the interval's half-width – do not intersect with each other.

The algorithms $\overline{\mathcal{A}}_U$ and $\underline{\mathcal{A}}_L$ are as follows:

- In both algorithms, 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 $2n+1$ zones $[x_{(i)}, x_{(i+1)}]$, where we denoted $x_{(0)} \stackrel{\text{def}}{=} -\infty$ and $x_{(2n+1)} \stackrel{\text{def}}{=} +\infty$.

- For each of 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 $x_{(i+1)} < \widetilde{x}_j - \frac{1+\alpha^2}{n} \cdot \Delta_j$, then we pick $x_j = \overline{x}_j$;
  - if $x_{(i+1)} > \widetilde{x}_j + \frac{1+\alpha^2}{n} \cdot \Delta_j$, 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 zone.
- To compute $\overline{U}$, for each of the sequences x_j, we check whether, for the selected values $x_1, \ldots, x_n$, the value of $E - \alpha \cdot \sigma$ is indeed within the corresponding zone, and if it is, compute the value $U = E + k_0 \cdot \sigma$. Finally, we return the largest of the computed values U as $\overline{U}$.
- To compute $\underline{L}$, for each of the sequences x_j, we check whether, for the selected values $x_1, \ldots, x_n$, the value of $E + \alpha \cdot \sigma$ is indeed within the corresponding zone, 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 4.2. *Let $1/n + 1/k_0^2 < 1$. The algorithms $\overline{\mathcal{A}}_U$ and $\underline{\mathcal{A}}_L$ compute $\overline{U}$ and $\underline{L}$ 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 4.3. *Let $1+(1/k_0)^2 < n$. For every positive integer C, the algorithms $\overline{\mathcal{A}}_U$ and $\underline{\mathcal{A}}_L$ compute $\overline{U}$ and $\underline{L}$ in quadratic time for all the cases in which no more than C "narrowed" intervals can have a common point.*

For each zone, we can determine the values of all optimal x_i – except for the case when the zone intersects with the corresponding 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 C different values i requires 2^C steps. Thus, 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 $r \stackrel{\text{def}}{=} |x - E|/\sigma$. 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.

Simplification 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 $x_i' \stackrel{\text{def}}{=} x_i - x$ 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 necessary and 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 $M \stackrel{\text{def}}{=} \dfrac{x_1^2 + \ldots + x_n^2}{n}$ is the second moment, we have $V/E^2 = M/E^2 - 1$, so computing the sharp bounds for V/E^2 is equivalent to computing the sharp bounds for the ratio $R \stackrel{\text{def}}{=} M/E^2$.

In this section, we will describe how to compute the sharp bounds $\underline{R}$ and $\overline{R}$ for the ratio R; based on these sharp bounds, we can compute the desired sharp 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 $\underline{R} \stackrel{\text{def}}{=} 1$. 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

$$e_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} \underline{x}_i + \sum_{j:\overline{x}_j \le x_{(k)}} \overline{x}_j,$$

$$m_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \geq x_{(k+1)}} (\underline{x}_i)^2 + \sum_{j:\overline{x}_j \leq x_{(k)}} (\overline{x}_j)^2,$$

and $n_k =$ the total number of such i's and j's. Then, we find $\lambda_k \stackrel{\text{def}}{=} m_k/e_k$. 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 $R_k \stackrel{\text{def}}{=} M_k/E_k^2$.

- 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 $\overline{R}$. In principle, we can have $\overline{R} = +\infty$ – e.g., if $0 \in [\underline{E}, \overline{E}]$. If $0 \notin [\underline{E}, \overline{E}]$ – e.g., if $\underline{E} > 0$ – then we can guarantee that $\overline{R} < +\infty$. In this case, we can bound $\overline{R}$ by the ratio $\overline{M}/\underline{E}^2$.

When $\overline{R} < n$, the maximum $\overline{R}$ is always attained at the endpoints:

Theorem 5.2. *When $\overline{R} < n$, 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 $x_i = \overline{x}_i$.*

In this case, 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 $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(2n)}$, 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 $x_i^+ \leq x_{(k)}$;
  - $x_i = \overline{x}_i$ if $x_i^- \geq x_{(k+1)}$;
  - both values $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$ otherwise.

For each of these combinations, we compute E, M, and $\lambda = M/E$, and check if λ is within the zone; if it is, we compute $R_k = M/E^2$.

The largest of these computed values 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 $\overline{R} < n$ and no more than C "narrowed" intervals can have a common point.*

6 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 $L \stackrel{\text{def}}{=} E - k_0 \cdot \sigma$ and $U \stackrel{\text{def}}{=} E + k_0 \cdot \sigma$ – 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 [4, 5], 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.

7 Proofs

Proof of Theorem 2.1. We will only prove the result for $\underline{U}$; for $\overline{L}$, the proof is practically identical.

Our proof is based on the fact that the minimum of a differentiable function of x_i on an interval $[\underline{x}_i, \overline{x}_i]$ is attained either inside this interval or at one of the endpoints. If the minimum is attained inside, the derivative $\dfrac{\partial U}{\partial x_i}$ is equal to 0; if it is attained at $x_i = \underline{x}_i$, then $\dfrac{\partial U}{\partial x_i} \geq 0$; finally, if it is attained at $x_i = \overline{x}_i$, then $\dfrac{\partial U}{\partial x_i} \leq 0$. For our function,

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

thus, $\dfrac{\partial U}{\partial x_i} = 0$ if and only if $x_i = \mu \stackrel{\text{def}}{=} E - \alpha \cdot \sigma$; similarly, the non-positiveness and non-negativeness of the derivative can be described by comparing x_i with μ. Thus:

- either $x_i \in (\underline{x}_i, \overline{x}_i)$ and $x_i = \mu$,
- or $x_i = \underline{x}_i$ and $x_i = \underline{x}_i \geq \mu$,
- or $x_i = \overline{x}_i$ and $x_i = \overline{x}_i \leq \mu$.

Hence, if we know how the value μ 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 cannot be attained inside or at the lower endpoint, so it is attained when $x_i = \overline{x}_i$;
- if $\mu \leq \underline{x}_i$, then, similarly, the minimum is attained when $x_i = \underline{x}_i$;
- if $\underline{x}_i < \mu < \overline{x}_i$, then the 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 zones.

Let the corresponding zone $[x_{(k)}, x_{(k+1)}]$ 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 the same value μ. To determine this μ, we will use the fact that, by definition, $\mu = E - \alpha \cdot \sigma$, where E and σ are computed by using the same value of μ. This equation is equivalent to $E - \mu \geq 0$ and $\alpha^2 \cdot \sigma^2 = (\mu - E)^2$. Substituting the above values of x_i into the formula for the mean E and for the standard deviation σ, we get the quadratic equation for μ which is described in the algorithm. So, for

each zone, we can uniquely determine the values x_i that may correspond to a minimum of U.

For the actual minimum, the value μ is inside one of these zone, so the smallest of the values U_k is indeed the desired minimum.

In this algorithm, sorting requires $O(n \cdot \log(n))$ steps (see, e.g., [1]), and the rest of the algorithm requires linear time ($O(n)$) for each of $2n+1$ zones, 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 $J \stackrel{\text{def}}{=} \sigma + \alpha \cdot E$ 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 polynomial time. 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 $\sum_{i=1}^{m} \eta_i \cdot s_i$ equals 0.

We will show that this problem can be reduced to the problem of computing $\overline{J}$ in polynomial time, 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 $M \stackrel{\text{def}}{=} (1/n) \cdot \sum_{i=1}^{n} x_i^2$ 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 $\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} \eta_i \cdot s_i = 0$. 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 $a \stackrel{\text{def}}{=} \frac{1}{m} \cdot \sum_{i=1}^{m} x_i = 0$. 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 Theorem 4.1. We will only prove the result for U; for L, the proof is practically identical.

When a function U attains its largest possible value at the value x_i inside the interval $[\underline{x}_i, \overline{x}_i]$, then at this inside point, $\dfrac{\partial U}{\partial x_i} = 0$ and $\dfrac{\partial^2 U}{\partial x_i^2} \le 0$. For our function U, we have

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

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

Since $\dfrac{\partial U}{\partial x_i} = 0$, we have $x_i - E = -\alpha \cdot \sigma$, hence

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

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 4.2–4.3. Similarly to the case of the previous two theorems, we will only provide the result for U; for L, the proof 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 ($x_j' = x_j$) 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 $U \ge U'$, where $U = E + k_0 \cdot \sigma$ and $U' = E' + k_0 \cdot \sigma'$. We can represent this inequality as $k_0 \cdot \sigma \ge (E' - E) + k_0 \cdot \sigma'$, hence either $(E' - E) + k_0 \cdot \sigma' \le 0$, or $k_0^2 \cdot \sigma^2 \ge (E' - E)^2 + k_0^2 \cdot (\sigma')^2 + 2(E - E') \cdot k_0 \cdot \sigma'$. In the second case, we move the terms linear in σ' to one side of the inequality and square both sides again. As a result, we get an inequality that only contains variances $V = \sigma^2 = M - E^2$ (where M is the sample second moment) and $V' = (\sigma')^2 = M' - (E')^2$ and no longer contains square roots.

For our choice of x_i', we have $E' = E - (2 \cdot \Delta_i)/n$ and

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

Substituting these expressions into the above-described inequality and simplifying the resulting algebraic expression, 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. This property leads to the inequality

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

So:

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

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

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

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 $\leq C$ values requires $\leq 2^C$ steps. For each zone and for each of these combinations, we need a linear time ($O(n)$) to compute 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.

Proof of Theorem 5.1. 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, $M \geq E^2$ and $R \geq 1$; hence $\underline{R} \geq 1$. If all the intervals have a common point, 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. For this case, the proof is similar to the proof of Theorem 2.1. Indeed, the minimum of a differentiable function of x_i on an interval $[\underline{x}_i, \overline{x}_i]$ is attained either inside this interval or at one of the endpoints. If the minimum is attained inside, the derivative $\dfrac{\partial R}{\partial x_i}$ is equal to 0; if it is attained at $x_i = \underline{x}_i$, then $\dfrac{\partial R}{\partial x_i} \geq 0$; finally, if it is attained at $x_i = \overline{x}_i$, then $\dfrac{\partial R}{\partial x_i} \leq 0$. For our function,

$$\frac{\partial R}{\partial x_i} = \frac{2}{n \cdot E^3} \cdot (E \cdot x_i - M);$$

thus, $\dfrac{\partial R}{\partial x_i} = 0$ if and only if $x_i = \lambda \stackrel{\text{def}}{=} M/E$; similarly, the non-positiveness and non-negativeness of the derivative can be described by comparing x_i with λ. Thus:

- either $x_i \in (\underline{x}_i, \overline{x}_i)$ and $x_i = \lambda$,
- or $x_i = \underline{x}_i$ and $x_i = \underline{x}_i \geq \lambda$,
- or $x_i = \overline{x}_i$ and $x_i = \overline{x}_i \leq \lambda$.

The proof continues just like for Theorem 2.1.

Proof of Theorem 5.2. This proof is similar to the proof of Theorem 4.1. When a function $R = M/E^2$ attains its largest possible value $\overline{R}$ at the value x_i inside the interval $[\underline{x}_i, \overline{x}_i]$, then at this inside point, $\dfrac{\partial R}{\partial x_i} = 0$ and $\dfrac{\partial^2 R}{\partial x_i^2} \leq 0$. For our function R, we have

$$\frac{\partial R}{\partial x_i} = \frac{2}{n \cdot E^3} \cdot (E \cdot x_i - M);$$

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

Since $\dfrac{\partial R}{\partial x_i} = 0$, we have $x_i = M/E$, hence

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

Since we assumed that $\overline{R} < n$, we conclude that the second derivative is positive and therefore, we cannot have a maximum in an internal point. The theorem is proven.

Proof of Theorem 5.3. This proof is similar to the proof of Theorems 4.2–4.3. Let us first prove that the algorithm described in Section 5 is indeed correct. Since $\overline{R}$, we can use Theorem 5.2 and conclude that the maximum of the function R 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 ($x'_j = x_j$) but replace x_i by $x'_i = \underline{x}_i = x_i - 2 \cdot \Delta_i$, then the value $R = M/E^2$ will decrease. We will denote the values of E and M that correspond to $(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)$, by E', and M'. In these terms, the desired inequality takes the form $M/E^2 \geq M'/(E')^2$, i.e., equivalently, $M \cdot (E')^2 \geq M' \cdot E^2$.

In the proof of Theorems 4.2–4.3, we had expressions for E' and M'. Substituting these expressions into the above inequality and simplifying the resulting algebraic expression, we conclude that

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

where $\lambda \stackrel{\text{def}}{=} M/E$.

Similarly, if the maximum is attained for $x_i = \underline{x}_i$, we have

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

Therefore, if we know the value of $\lambda = M/E$, then:

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

Similarly to the proof of Theorems 4.2–4.3, we can now conclude that the algorithm from Section 5 is correct and that this algorithm requires quadratic time.

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).

The authors are thankful to the anonymous referees for valuable suggestions.

References

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, and Mc-Graw Hill Co., N.Y., 2001.

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

[3] 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.

[4] 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.

[5] 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.

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

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

[8] O. Kosheleva, S. Cabrera, R. Osegueda, S. Nazarian, D. L. George, M. J. George, V. Kreinovich, and K. Worden, "Case study of nonlinear 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.

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

[10] V. Kreinovich, L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg, "Outlier Detection Under Interval Uncertainty: Algorithmic Solvability and Computational Complexity", In: I. Lirkov, S. Margenov, J. Wasniewski, and P. Yalamov (eds.), *Large-Scale Scientific Computing, Proceedings of the 4-th International Conference LSSC'2003, Sozopol, Bulgaria, June 4–8, 2003*, Springer Lecture Notes in Computer Science, 2004, Vol. 2907, pp. 238–245.

[11] 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, pp. 401–406.

[12] 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

[13] 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.

[14] 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.

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

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

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.