Journal of Uncertain Systems
Vol.9, No.2, pp.124-132, 2015
Online at: www.jus.org.uk

Computing Standard-Deviation-to-Mean and Variance-to-Mean Ratios under Interval Uncertainty is NP-Hard

Sio-Long Lo*
Faculty of Information Technology, Macau University of Science and Technology (MUST)
Avenida Wai Long, Taipa, Macau SAR, China

Received 15 January 2014; Revised 16 February 2014

Abstract

Once we have a collection of values corresponding a class of objects, a usual way to decide whether a new object with the value of the corresponding property belongs to this class is to check whether this value belongs to interval from mean E minus k sigma σ to mean plus k sigma, where the parameter k is determined by the degree of confidence with which we want to make the decision. For each value x, the degree of confidence that x belongs to the class depends on the smallest value k for which x belongs to the corresponding interval, i.e., on the ratio r of σ and $|E-x|$. In practice, we often only know the intervals that contain the actual values. Different values from these intervals lead, in general, to different values of r, so it is desirable to compute the range of corresponding values of r. Polynomial-time algorithms are known for computing this range under certain conditions; whether it is always possible to compute this range in polynomial time was unknown. In this paper, we prove that the problem of computing this range is NP-hard. A similar NP-hardness result is proven for a similar ratio between the variance V and the mean E which is used in clustering.

Keywords: statistics under interval uncertainty, variance-to-mean ratio, standard-deviation-to-mean ratio

1 Formulation of the Problem

A practical problem: checking whether an object belongs to a class. In many practical situations, we want to check whether a new object belongs to a given class. In such situations, we usually have a sample of objects which are known to belong to this class. For example, a biologists who is studying bats has observed several bats from a local species; the question is whether a newly observed bat belongs to the same species – or to a different bat species.

To solve this problem, we usually measure one or more quantities for the objects from this class and for the new object, and compare the resulting values. For the simplest case of a single quantity, we have a collection of values $x_1, \ldots, x_n$ corresponding to objects from the known class, and a value x corresponding to the new object.

A standard way to decide whether an object belongs to a class. A usual way to decide whether a new object with the value x belongs to the class characterized by the values $x_1, \ldots, x_n$ is to check whether the value x belongs to the "k sigma" interval $[E-k\cdot\sigma, E+k\cdot\sigma]$, where:

- $E \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} x_i/n$ is the sample mean,
- $\sigma = \sqrt{V}$, where $V \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} (x_i-E)^2/n$ is the sample variance, and

*Corresponding author.
Email: akennetha@gmail.com (S.-L. Lo).

- the parameter k is determined by the degree of confidence with which we want to make the decision; usually, we take $k = 2$ (corresponding to confidence 0.9), $k = 3$ (corresponding to 0.999), or $k = 6$ (corresponding to $1 - 10^{-8}$);

see, e.g., [7, 8].

How confident are we about this decision? For each value x, when k is large enough, the value x belongs to the interval $[E - k \cdot \sigma, E + k \cdot \sigma]$. Our degree of confidence that x belongs to the class depends on the smallest values k^- for which $x \geq E - k^- \cdot \sigma$ and on the smallest value k^+ for which $x \leq E + k^+ \cdot \sigma$. For example, if one of the values k^- and k^+ is larger than 2, then our confidence is smaller than $1 - 0.9 = 10\%$; if one of these values exceeds 3, our confidence is $\leq 0.1\%$, etc.

How to compute the parameters describing confidence? The inequality $x \geq E - k^- \cdot \sigma$ is equivalent to $k^- \cdot \sigma \geq E - x$ and $k^- \geq (E - x)/\sigma$. Thus, when $x < E$, the corresponding smallest value is equal to $k^- = (E - x)/\sigma$.

Similarly, the inequality $x \leq E + k^+ \cdot \sigma$ is equivalent to $k^+ \cdot \sigma \geq x - E$ and $k^+ \geq (x - E)/\sigma$. Thus, when $x > E$, the corresponding smallest value is equal to $k^+ = (x - E)/\sigma$.

So, to determine the parameter describing confidence, we must compute one of the ratios $k^- \stackrel{\text{def}}{=} (E - x)/\sigma$ or $k^+ \stackrel{\text{def}}{=} (x - E)/\sigma$. Often, reciprocal ratio are used:

$$r^- \stackrel{\text{def}}{=} \frac{1}{k^-} = \frac{\sigma}{E - x}$$

and

$$r^+ \stackrel{\text{def}}{=} \frac{1}{k^+} = \frac{\sigma}{x - E}.$$

Case of interval uncertainty. The traditional formulas are based on the simplifying assumptions that we know the exact values $x_1, \ldots, x_n$ of the corresponding quantity. In practice, these values come from measurement, and measurements are never absolutely accurate; see, e.g. [7]. It is therefore necessary to take this measurement uncertainty into account when computing the corresponding ratios. In other words, it is necessary to take into account that the measured values $\widetilde{x}_1, \ldots, \widetilde{x}_n$ are, in general, different from the actual (unknown) values $x_1, \ldots, x_n$.

Traditional engineering techniques for taking uncertainty into account assume that we know the probabilities of different values of measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. In many practical situations, however, we only know the upper bound Δ_i on this measurement error, i.e., the value for which $|\Delta x_i| \leq \Delta_i$; see, e.g., [7]. In this case, once we know the measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$, the only information that we have about each actual value x_i is that this value belongs to the interval $\mathbf{x}_i \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

Different possible values $x_i \in \mathbf{x}_i$ lead, in general, to different values of the corresponding ratios $r(x_1, \cdots, x_n)$. Thus, it is desirable to compute the *range* of possible values of this ratio:

$$\mathbf{r} = [\underline{r}, \overline{r}] \stackrel{\text{def}}{=} \{r(x_1, \cdots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}. \tag{1}$$

Comment. This problem is a particular case of a problem of computing the range of a function under interval uncertainty, the problem known as *interval computation*; see, e.g., [3, 5].

What is known. The problem of computing the range (1) was analyzed in [4] – together with similar problems of computing ranges for the thresholds $E - k \cdot \sigma$ and $E + k \cdot \sigma$ for a given k; see also [1]. In these papers, feasible algorithms are described for computing the upper bounds for $E - k \cdot \sigma$, and for computing the lower bounds for $E + k \cdot \sigma$, $\sigma/(E - x)$, and $\sigma/(x - E)$.

Algorithms are also described for computing the remaining bounds under certain conditions on the intervals: namely, for computing the lower bounds for $E - k \cdot \sigma$, and for computing the upper bounds for $E + k \cdot \sigma$, $\sigma/(E - x)$, and $\sigma/(x - E)$. Such conditions are necessary: in [4], it is proven that, in general, the problems of computing the lower bound for $E - k \cdot \sigma$ and the upper bounds for $E + k \cdot \sigma$ are NP-hard – which means that, unless P=NP, these problems cannot be, in general, solved in polynomial (= feasible) time; see, e.g., [2, 6].

What we do in this paper. While it was known that computing bounds for the thresholds $E - k \cdot \sigma$ and $E + k \cdot \sigma$, whether the problem for computing the range of the ratio is NP-hard was not known. In this paper, we prove that this problem is also NP-hard.

We use the same idea to prove the NP-hardness of a similar problem: of computing the range of a ratio V/E used in clustering.

2 Results

Discussion. In order to prove that a problem is NP-hard, it is sufficient to prove that a particular case of this problem is NP-hard. Thus, to prove that the general problem of computing the upper bound of the ratios $\sigma/(E - x)$ and $\sigma/(x - E)$ is NP-hard, it is sufficient to prove that computing the range of the standard-deviation-to-mean ratio σ/E (corresponding to $x = 0$) is NP-hard. Moreover, it is sufficient to prove this for the case when all the intervals $[\underline{x}_i, \overline{x}_i]$ contain only non-negative values, i.e., when $\underline{x}_i \geq 0$ for all i.

Theorem 1. *The following problem is NP-hard:*

- given: *a natural number n and n (rational-valued) intervals $[\underline{x}_i, \overline{x}_i]$,*
- compute: *the upper endpoint $\overline{r}$ of the range*

$$\mathbf{r} = [\underline{r}, \overline{r}] = \{r(x_1, \cdots, x_n) \,|\, x_1 \in [\underline{x}_1, \overline{x}_1], \ldots, x_n \in [\underline{x}_n, \overline{x}_n]\}$$

of the ratio $r = \sqrt{V}/E$, where $E = \sum\limits_{i=1}^{n} x_i/n$ and $V = \sum\limits_{i=1}^{n} (x_i - E)^2/n$.

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

Theorem 2. *The following problem is NP-hard:*

- given: *a natural number n and n (rational-valued) intervals $[\underline{x}_i, \overline{x}_i]$,*
- compute: *the upper endpoint $\overline{r}$ of the range*

$$\mathbf{r} = [\underline{r}, \overline{r}] = \{r(x_1, \cdots, x_n) \,|\, x_1 \in [\underline{x}_1, \overline{x}_1], \ldots, x_n \in [\underline{x}_n, \overline{x}_n]\}$$

of the ratio $r = V/E$, where $E = \sum\limits_{i=1}^{n} x_i/n$ and $V = \sum\limits_{i=1}^{n} (x_i - E)^2/n$.

3 Proofs

3.1 Proof of Theorem 1

1°. The above expression for the ratio r uses a square root – to compute $\sigma = \sqrt{V}$. In optimization, we usually use derivatives, and the square root function $f(x) = \sqrt{x}$ has infinite derivative when $x = 0$. To avoid this problem, we can use the fact that $r = \sqrt{R}$, where $R \stackrel{\text{def}}{=} V/E^2$, and that the function $\sqrt{x}$ is strictly increasing. Thus,

- the smallest possible value $\underline{r}$ of r is equal to the square root of the smallest possible value of R: $\underline{r} = \sqrt{\underline{R}}$; and
- the largest possible value $\overline{r}$ of r is equal to the square root of the largest possible value of R: $\overline{r} = \sqrt{\overline{R}}$.

Thus, the problem of computing the range of the ratio r is feasibly equivalent to the problem of computing the range $[\underline{R}, \overline{R}]$ of the new ratio R. In particular, this means that to prove NP-hardness of the original range computation problem, it is sufficient to prove that the new range computation problem is NP-hard.

2°. Similarly to the NP-hardness proofs for the thresholds [4], to prove the NP-hardness of our problem, we will show that a known NP-hard problem – the subset sum problem – can be reduced to it. In this problem, we are given n positive integers $s_1, \ldots, s_n$, and we need to check whether there exists signs $\eta_i \in \{-1, 1\}$ for which $\sum\limits_{i=1}^{n} \eta_i \cdot s_i = 0$.

Specifically, we will prove that such signs exist if and only if for an appropriately chosen integer N and for the intervals $\mathbf{x}_i = [N - s_i, N + s_i]$, the upper endpoint $\overline{R}$ of the range $[\underline{R}, \overline{R}]$ of the variance-to-squared-mean ratio R is greater than or equal to $R_0 \stackrel{\text{def}}{=} M_0/N^2$, where $M_0 \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} s_i^2/n$.

Comment. Such a reduction is a standard way of proving NP-hardness. Indeed, by definition, a problem is NP-hard if every problem from a certain class NP can be reduced to it [2, 6]. Thus, if a known NP-hard problem $\mathcal{P}$ can be reduced to a given problem $\mathcal{P}_0$, then, since every problem from the class NP can be reduced to $\mathcal{P}$ and $\mathcal{P}$ can be reduced to $\mathcal{P}_0$, every problem from the class NP can also be reduced to $\mathcal{P}_0$ – and thus, our problem $\mathcal{P}_0$ is indeed NP-hard.

3°. Let us prove that the ratio $R = V/E^2$ attains its maximum on the box $[\underline{x}_1, \overline{x}_1] \times \cdots \times [\underline{x}_n, \overline{x}_n]$ when each of the variables x_i is equal to one of the endpoints $\underline{x}_i$ or $\overline{x}_i$.

We will prove this statement by contradiction. Let us assume that for some i, the function $R(x_1, \ldots, x_n)$ attains its maximum on an interval $[\underline{x}_i, \overline{x}_i]$ at an internal point $x_i \in (\underline{x}_i, \overline{x}_i)$. In this case, according to calculus, at this point, the partial derivative $\dfrac{\partial R}{\partial x_i}$ should be equal to 0, and the second derivative $\dfrac{\partial^2 R}{\partial x_i^2}$ should be non-positive.

Here,

$$\frac{\partial E}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{1}{n} \cdot \sum_{j=1}^{n} x_j\right) = \frac{1}{n} \tag{2}$$

and, since $V = M - E^2$, where $M \stackrel{\text{def}}{=} \sum\limits_{j=1}^{n} x_j^2/n$, we have

$$\frac{\partial V}{\partial x_i} = \frac{\partial M}{\partial x_i} - \frac{\partial E^2}{\partial x_i}. \tag{3}$$

Here,

$$\frac{\partial E^2}{\partial x_i} = 2 \cdot E \cdot \frac{\partial E}{\partial x_i} = 2 \cdot E \cdot \frac{1}{n}. \tag{4}$$

Since

$$\frac{\partial M}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{1}{n} \cdot \sum_{j=1}^{n} x_j^2\right) = \frac{1}{n} \cdot 2x_i, \tag{5}$$

we have

$$\frac{\partial V}{\partial x_i} = \frac{1}{n} \cdot 2x_i - 2 \cdot E \cdot \frac{1}{n}. \tag{6}$$

Thus,

$$\frac{\partial R}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{V}{E^2}\right) = \frac{\dfrac{\partial V}{\partial x_i} \cdot E^2 - V \cdot \dfrac{\partial E^2}{\partial x_i}}{E^4}$$

$$= \frac{\left(\dfrac{1}{n} \cdot 2x_i - 2 \cdot E \cdot \dfrac{1}{n}\right) \cdot E^2 - V \cdot 2 \cdot E \cdot \dfrac{1}{n}}{E^4} = 2 \cdot \frac{x_i \cdot E - E^2 - V}{n \cdot E^3}. \tag{7}$$

Thus, when $\dfrac{\partial R}{\partial x_i} = 0$, we get

$$x_i = \frac{E^2 + V}{E} = \frac{M}{E}. \tag{8}$$

Differentiating $\frac{\partial R}{\partial x_i}$ with respect to x_i, and using the expressions (2) and (6), we get the following expression for the second derivative:

$$\frac{\partial^2 R}{\partial x_i^2} = 2 \cdot \frac{3 \cdot V + (3+n) \cdot E^2 - 4 \cdot x_i \cdot E}{n^2 \cdot E^4}. \tag{9}$$

The denominator is positive, so since the second derivative is non-positive, we conclude that the numerator must be non-positive as well, i.e., that

$$\begin{aligned} & 3 \cdot V + (3+n) \cdot E^2 - 4 \cdot x_i \cdot E \\ = \; & 3 \cdot (M - E^2) + (3+n) \cdot E^2 - 4 \cdot x_i \cdot E \\ = \; & 3 \cdot M + n \cdot E^2 - 4 \cdot x_i \cdot E \\ \le \; & 0. \end{aligned} \tag{10}$$

By definition,

$$E = \frac{1}{n} \cdot \sum_{j=1}^{n} x_i = \frac{1}{n} \cdot x_i + \frac{1}{n} \cdot E_i, \tag{11}$$

where we denoted

$$E_i \stackrel{\text{def}}{=} \sum_{j \ne i} x_j. \tag{12}$$

Similarly,

$$M = \frac{1}{n} \cdot x_i^2 + \frac{1}{n} \cdot M_i, \tag{13}$$

where

$$M_i \stackrel{\text{def}}{=} \sum_{j \ne i} x_j^2. \tag{14}$$

Substituting the formulas (11) and (13) into the right-hand side of the inequality (10), we conclude that

$$3 \cdot \left(\frac{1}{n} \cdot x_i^2 + \frac{1}{n} \cdot M_i \right) + n \cdot \left(\frac{1}{n} \cdot x_i + \frac{1}{n} \cdot E_i \right)^2 - 4 \cdot x_i \cdot \left(\frac{1}{n} \cdot x_i + \frac{1}{n} \cdot E_i \right) \le 0. \tag{15}$$

Multiplying both sides of this inequality by n, we get

$$3 \cdot (x_i^2 + M_i) + (x_i + E_i)^2 - 4 \cdot x_i \cdot (x_i + E_i) \le 0. \tag{16}$$

Opening parentheses, we get

$$3 \cdot x_i^2 + 3 \cdot M_i + x_i^2 + 2 \cdot x_i \cdot E_i + E_i^2 - 4 \cdot x_i^2 - 4 \cdot x_i \cdot E_i = 3 \cdot M_i + E_i^2 - 2 \cdot x_i \cdot E_i \le 0. \tag{17}$$

Here, due to (8), we have $x_i \cdot E = M$, hence, substituting expressions (11) and (13)

$$x_i \cdot \left(\frac{1}{n} \cdot x_i + \frac{1}{n} \cdot E_i \right) = \frac{1}{n} \cdot x_i^2 + \frac{1}{n} \cdot M_i. \tag{18}$$

Multiplying both sides of this equality by n and canceling equal terms x_i^2 in both sides, we get $x_i \cdot E_i = M_i$. Substituting M_i instead of $x_i \cdot E_i$ into the right-hand side of the inequality (17), we conclude that $M_i + E_i^2 \le 0$.

However, for large enough N (specifically, for $N > \max\limits_i s_i$), we have $x_j \ge N - s_j > 0$, hence $E_i = \sum\limits_{j \ne i} x_j > 0$, $M_i = \sum\limits_{j \ne i} x_j^2 > 0$, and thus, $M_i + E_i^2 > 0$. This contradiction shows that the maximum of the ratio R cannot be attained at an internal point of the interval $(\underline{x}_i, \overline{x}_i)$. Thus, this maximum can only be attained when $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$.

4°. Let us now prove that the maximum $\overline{R}$ is greater than or equal to $R_0 = M_0/N^2$ if and only if there exist signs $\eta_i \in \{-1, 1\}$ for which $\sum\limits_{i=1}^{n} \eta_i \cdot s_i = 0$.

4.1°. If such signs exist, then we can take $x_i = N + \eta_i \cdot s_i$. For these values, due to the properties of the signs, we have $E = N$ and therefore, $x_i - E = \pm s_i$ and

$$V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 = M_0,$$

and $R = V/E^2 = M_0/N^2$. The largest possible value $\overline{R}$ must therefore be larger than or equal to this value.

4.2°. Vice versa, let us assume that $\overline{R} \geq R_0$. Let x_i be the values for which the ratio R attains its maximum value $\overline{R}$.

Due to Part 3 of this proof, this maximum is attained when each variable x_i is equal to either $N - s_i$ or to $N + s_i$, i.e., when $x_i = N + t_i$ with $t_i = \eta_i \cdot s_i$. In this case, $E = N + e$, where $e \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} t_i/n$. Since the variance does not change if we simply shift all the values by N, we have

$$V(x_1, \cdots, x_n) = V(t_1, \cdots, t_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} t_i^2 - e^2.$$

Since $t_i = \pm s_i$, we have $t_i^2 = s_i^2$ and thus,

$$\frac{1}{n} \cdot \sum_{i=1}^{n} t_i^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 = M_0$$

and $V = M_0 - e^2$. Thus,

$$\overline{R} = \frac{V}{E^2} = \frac{M_0 - e^2}{(N+e)^2},$$

and the inequality $\overline{R} \geq R_0$ takes the form

$$\frac{M_0 - e^2}{(N+e)^2} \geq \frac{M_0}{N^2}. \tag{19}$$

Multiplying both sides by the common denominator $(N+e)^2 \cdot N^2$ and opening parentheses, we conclude that

$$N^2 \cdot M_0 - e^2 \cdot N^2 \geq M_0 \cdot N^2 + 2M_0 \cdot N \cdot e + M_0 \cdot e^2.$$

Canceling the term $M_0 \cdot N^2$ in both sides, and moving all the terms to the right-hand side, we get

$$e^2 \cdot (N^2 + M_0) + 2 \cdot M_0 \cdot N \cdot e \leq 0. \tag{20}$$

If $e > 0$, then the left-hand side is positive and cannot be ≤ 0, so $e \leq 0$. If $e < 0$, then (20) becomes

$$|e|^2 \cdot (N^2 + M_0) - 2 \cdot M_0 \cdot N \cdot |e| \leq 0. \tag{21}$$

Dividing both sides by $|e| > 0$, we get

$$|e| \cdot (N^2 + M_0) - 2 \cdot M_0 \cdot N \leq 0, \tag{22}$$

hence

$$|e| \leq \frac{2 \cdot M_0 \cdot N}{N^2 + M_0}. \tag{23}$$

When N increases, the right-hand side of this inequality tends to 0. However, by definition, all the values s_i are integers, so all the values $t_i = \pm s_i$ are also integers, the sum $n \cdot e = \sum\limits_{i=1}^{n} t_i$ is an integer. Since $e \neq 0$, the absolute value $|n \cdot e|$ of this integer must be at least 1, so $|n \cdot e| \geq 1$ and $|e| \geq 1/n$.

Since

$$\frac{2 \cdot M_0 \cdot N}{N^2 + M_0} \to 0$$

as $N \to \infty$, for sufficiently large N, we have

$$\frac{1}{n} > \frac{2 \cdot M_0 \cdot N}{N^2 + M_0}, \tag{24}$$

and thus, the inequality (23) is impossible. This shows that e cannot be negative, hence $e = 0$, and thus, $n \cdot e = \sum\limits_{i=1}^{n} \eta_i \cdot s_i = 0$. The theorem is proven.

3.2 Proof of Theorem 2

1°. We will reduce our problem to the same known NP-hard problem as in the proof of Theorem 1: given n integers $s_1, \ldots, s_n$, check whether there exists signs $\eta_i \in \{-1, 1\}$ for which $\sum\limits_{i=1}^{n} \eta_i \cdot s_i = 0$. Specifically, we will show that for a sufficiently large N, if we take $x_i \in [N - s_i, N + s_i]$, then the upper endpoint $\overline{r}$ of the ratio $r = V/E$ is greater than or equal to $r_0 \stackrel{\text{def}}{=} M_0/N$, where $M_0 \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} s_i^2/n$ if and only if there exists signs $\eta_i \in \{-1, 1\}$ for which $\sum\limits_{i=1}^{n} \eta_i \cdot s_i = 0$.

2°. Let us prove that the ratio $r = V/E$ attains its maximum on the box $[\underline{x}_1, \overline{x}_1] \times \cdots \times [\underline{x}_n, \overline{x}_n]$ when each of the variables x_i is equal to one of the endpoints $\underline{x}_i$ or $\overline{x}_i$.

Indeed, as in the proof of Theorem 1, if the maximum is attained inside an interval $(\underline{x}_i, \overline{x}_i)$, then we should have $\dfrac{\partial r}{\partial x_i} = 0$ and $\dfrac{\partial^2 r}{\partial x_i^2} \le 0$.

Here,

$$\frac{\partial r}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{V}{E}\right) = \frac{\dfrac{\partial V}{\partial x_i} \cdot E - V \cdot \dfrac{\partial E}{\partial x_i}}{E^2}. \tag{25}$$

Using the formulas (2) and (6), we conclude that

$$\frac{\partial r}{\partial x_i} = \frac{2 \cdot x_i \cdot E - 2 \cdot E^2 - V}{nE^2}. \tag{26}$$

Similarly, by using the same formulas (2) and (6), we conclude that

$$\frac{\partial^2 r}{\partial x_i^2} = \frac{\partial}{\partial x_i}\left(\frac{\partial r}{\partial x_i}\right) = 2 \cdot \frac{V - 2x_i \cdot E + (n+1) \cdot E^2}{n^2 E^3}. \tag{27}$$

Since $V = M - E^2$, we have

$$\frac{\partial^2 r}{\partial x_i^2} = 2 \cdot \frac{M - E^2 - 2x_i \cdot E + (n+1) \cdot E^2}{n^2 E^3} = 2 \cdot \frac{M - 2x_i \cdot E + n \cdot E^2}{n^2 E^3}. \tag{28}$$

Multiplying both the numerator and the denominator by n and taking into account that

$$n \cdot M^2 = \sum_{j=1}^{n} x_j^2 = x_i^2 + \sum_{j \ne i} x_j^2,$$

we conclude that

$$\frac{\partial^2 r}{\partial x_i^2} = 2 \cdot \frac{\sum\limits_{j \ne i} x_j^2 + (n \cdot E)^2 - 2 \cdot n \cdot E \cdot x_i + x_i^2}{n^3 \cdot E^3}. \tag{29}$$

The last three terms in the numerator form a full square, so

$$\frac{\partial^2 r}{\partial x_i^2} = 2 \cdot \frac{\sum\limits_{j \ne i} x_j^2 + (n \cdot E - x_i)^2}{n^3 \cdot E^3}. \tag{30}$$

When $N > \max(s_i)$, we have $x_i \ge N - s_i > 0$ hence $\dfrac{\partial^2 r}{\partial x_i^2} > 0$. Thus, the maximum cannot be attained at any internal point. The statement is proven.

3°. Let us now prove that the maximum $\overline{r}$ is greater than or equal to $r_0 = M_0/N$ if and only if there exist signs $\eta_i \in \{-1, 1\}$ for which $\sum\limits_{i=1}^{n} \eta_i \cdot s_i = 0$.

3.1°. If such signs exist, then we can take $x_i = N + \eta_i \cdot s_i$. For these values, due to the properties of the signs, we have $E = N$ and therefore, $x_i - E = \pm s_i$ and

$$V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 = M_0,$$

and $r = V/E = M_0/N$. The largest possible value $\overline{r}$ must therefore be larger than or equal to this value.

3.2°. Vice versa, let us assume that $\overline{r} \geq r_0$. Let x_i be the values for which the ratio r attains its maximum value $\overline{r}$.

Due to Part 2 of this proof, this maximum is attained when each variable x_i is equal to either $N - s_i$ or to $N + s_i$, i.e., when $x_i = N + t_i$ with $t_i = \eta_i \cdot s_i$. In this case, $E = N + e$, where $e \stackrel{\text{def}}{=} \sum_{i=1}^{n} t_i/n$. Since the variance does not change if we simply shift all the values by N, we have

$$V(x_1, \cdots, x_n) = V(t_1, \cdots, t_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} t_i^2 - e^2.$$

Since $t_i = \pm s_i$, we have $t_i^2 = s_i^2$ and thus,

$$\frac{1}{n} \cdot \sum_{i=1}^{n} t_i^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 = M_0$$

and $V = M_0 - e^2$. Thus,

$$\overline{r} = \frac{V}{E} = \frac{M_0 - e^2}{N + e},$$

and the inequality $\overline{r} \geq R_0$ takes the form

$$\frac{M_0 - e^2}{N + e} \geq \frac{M_0}{N}. \tag{31}$$

Multiplying both sides by the common denominator $(N + e) \cdot N$ and opening parentheses, we conclude that

$$N \cdot M_0 - e^2 \cdot N \geq M_0 \cdot N + M_0 \cdot e.$$

Canceling the term $M_0 \cdot N$ in both sides, and moving all the terms to the right-hand side, we get

$$e^2 \cdot N + M_0 \cdot e \leq 0. \tag{32}$$

If $e > 0$, then the left-hand side is positive and cannot be ≤ 0, so $e \leq 0$. If $e < 0$, then (32) becomes

$$|e|^2 \cdot N - M_0 \cdot |e| \leq 0. \tag{33}$$

Dividing both sides by $|e| > 0$, we get

$$|e| \cdot N - M_0 \leq 0, \tag{34}$$

hence

$$|e| \leq \frac{M_0}{N}. \tag{35}$$

When N increases, the right-hand side of this inequality tends to 0. However, as we have mentioned in the proof of Theorem 1, when $e \neq 0$, we have $|e| \geq 1/n$.

Since $M_0/N \to 0$ as $N \to \infty$, for sufficiently large N, we have

$$\frac{1}{n} > \frac{M_0}{N}, \tag{36}$$

and thus, the inequality (35) is impossible. This shows that e cannot be negative, hence $e = 0$, and thus, $n \cdot e = \sum_{i=1}^{n} \eta_i \cdot s_i = 0$. The theorem is proven.

Acknowledgments

This project was done during Sio-Long Lo's visit to the University of Texas at El Paso. Sio-Long is thankful to the Macau University of Science and Technology (MUST) and to the University of Texas at El Paso (UTEP) for this research opportunity, and to Professors Liya Ding (MUST) and Vladik Kreinovich (UTEP) for their support.

References

[1] Dantsin, E., Wolpert, A., Ceberio, M., Xiang, G., and V. Kreinovich, Detecting outliers under interval uncertainty: a new algorithm based on constraint satisfaction, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp.802–809, 2006.

[2] Garey, M.R., and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. F. Freeman, San Francisco, 1979.

[3] Jaulin, L., Kieffer, M., Didrit, O., and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics*, Springer-Verlag, London, 2001.

[4] Kreinovich, V., Longpré, L., Patangay, P., Ferson, S., and L. Ginzburg, Outlier detection under interval uncertainty: algorithmic solvability and computational complexity, *Reliable Computing*, vol.11, no.1, pp.59–76, 2005.

[5] Moore, R.E., Kearfott, R.B., and M.J. Cloud, *Introduction to Interval Analysis*, SIAM Press, Philadelphia, Pennsylviania, 2009.

[6] Papadimitriou, C., *Computational Complexity*, Addison Welsey, Reading, Massachusetts, 1994.

[7] Rabinovich, S., *Measurement Errors and Uncertainties: Theory and Practice*, American Institute of Physics, New York, 2005.

[8] Sheskin, D.J., *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.