Computing Variance for Interval Data is NP-Hard*

Scott Ferson[1], Lev Ginzburg[1],
Vladik Kreinovich[2], Luc Longpré[2], and Monica Aviles[2]

[1]Applied Biomathematics, 100 North Country Road,
Setauket, NY 11733, USA, {scott,lev}@ramas.com
[2]Computer Science Department, University of Texas at El Paso
El Paso, TX 79968, USA, {maviles,longpre,vladik}@cs.utep.edu

Abstract

When we have only interval ranges $[\underline{x}_i, \overline{x}_i]$ of sample values $x_1, \ldots, x_n$, what is the interval $[\underline{V}, \overline{V}]$ of possible values for the variance V of these values? We prove that the problem of computing the upper bound $\overline{V}$ is NP-hard. We provide a feasible (quadratic time) algorithm for computing the lower bound $\underline{V}$ on the variance of interval data. We also provide a feasible algorithm that computes $\overline{V}$ under reasonable easily verifiable conditions.

1 Introduction

When we have n measurement results $x_1, \ldots, x_n$, traditional statistical approach usually starts with computing their (sample) average $E = \bar{x} = \frac{x_1 + \ldots + x_n}{n}$ and their (sample) variance $V = \frac{(x_1 - E)^2 + \ldots + (x_n - E)^2}{n-1}$.

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 $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

As a result, the sets of possible values of E and V are also intervals. Since E is an increasing function of each of the variables x_i, it is easy to compute the interval $\mathbf{E} = [\underline{E}, \overline{E}]$ of possible values of E: $\underline{E} = \frac{\underline{x}_1 + \ldots + \underline{x}_n}{n}$; $\overline{E} = \frac{\overline{x}_1 + \ldots + \overline{x}_n}{n}$. What is the interval $[\underline{V}, \overline{V}]$ of possible values for sample variance V?

When all the intervals $\mathbf{x}_i$ intersect, then it is possible that all the actual (unknown) values $x_i \in \mathbf{x}_i$ are the same and hence, that the sample variance is 0. In other words, if the intervals have a non-empty intersection, then $\underline{V} = 0$. Conversely, if the intersection of $\mathbf{x}_i$ is empty, then V cannot be 0, hence $\underline{V} > 0$. The question is (see, e.g., [4]): What is the total set of possible values of V when the above intersection is empty?

2 First Result: Computing $\overline{V}$ is NP-Hard

Theorem 2.1. *Computing $\overline{V}$ is NP-hard.*

*Copyright ©Scott Ferson, Lev Ginzburg, Vladik Kreinovich, Luc Longpré, and Monica Aviles, 2002

The very fact that computing the range of a quadratic function is NP-hard was first proven by Vavasis [3] (see also [1]). We have shown that this difficulty happens even for very simple quadratic functions V frequently used in data processing.

A natural question is: maybe the difficulty comes from the requirement that the range be computed exactly? In practice, it is often sufficient to compute, in a reasonable amount of time, a usefully accurate estimate $\widetilde{\overline{V}}$ for $\overline{V}$, i.e., an estimate $\widetilde{\overline{V}}$ which is accurate with a given accuracy $\varepsilon > 0$: $\left|\widetilde{\overline{V}} - \overline{V}\right| \le \varepsilon$. Alas, for any ε, such computations are also NP-hard:

Theorem 2.2. *For every $\varepsilon > 0$, the problem of computing $\overline{V}$ with accuracy ε is NP-hard.*

3 Second Result: Computing $\underline{V}$

First, we design a *feasible* algorithm for computing the exact lower bound $\underline{V}$ of the sample variance. Specifically, our algorithm is *quadratic-time*, i.e., it requires $O(n^2)$ computational steps (arithmetic operations or comparisons) for n interval data points $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.

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

- First, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}$.
- Second, we compute $\underline{E}$ and $\overline{E}$ and select all "small intervals" $[x_{(k)}, x_{(k+1)}]$ that intersect with $[\underline{E}, \overline{E}]$.
- For each of selected small intervals $[x_{(k)}, x_{(k+1)}]$, we compute the ratio $r_k = S_k/N_k$, where

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

  and N_k is the total number of such i's and j's If $r_k \in [x_{(k)}, x_{(k+1)}]$, then we compute

$$V_k' \stackrel{\text{def}}{=} \frac{1}{n-1} \cdot \left(\sum_{i:\underline{x}_i > x_{(k+1)}} (\underline{x}_i - r)^2 + \sum_{j:\overline{x}_j < x_{(k)}} (\overline{x}_j - r)^2 \right).$$

  If $N_k = 0$, we take $V_k' \stackrel{\text{def}}{=} 0$.
- Finally, we return the smallest of the values V_k' as $\underline{V}$.

Theorem 3.1. *The algorithm $\underline{\mathcal{A}}$ always compute $\underline{V}$ is quadratic time.*

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

We have implemented this algorithm in C++, it works really fast.

Example. We start with 5 intervals: $\mathbf{x}_1 = [2.1, 2.6]$, $\mathbf{x}_2 = [2.0, 2.1]$, $\mathbf{x}_3 = [2.2, 2.9]$, $\mathbf{x}_4 = [2.5, 2.7]$, and $\mathbf{x}_5 = [2.4, 2.8]$. After sorting the bounds, we get the following "small intervals": $[x_{(1)}, x_{(2)}] = [2.0, 2.1]$, $[x_{(2)}, x_{(3)}] = [2.1, 2.1]$, $[x_{(3)}, x_{(4)}] = [2.1, 2.2]$, $[x_{(4)}, x_{(5)}] = [2.2, 2.4]$, $[x_{(5)}, x_{(6)}] = [2.4, 2.5]$, $[x_{(6)}, x_{(7)}] = [2.5, 2.6]$, $[x_{(7)}, x_{(8)}] = [2.6, 2.7]$, $[x_{(8)}, x_{(9)}] = [2.7, 2.8]$, and $[x_{(9)}, x_{(10)}] = [2.8, 2.9]$.

The interval for sample average is $\mathbf{E} = [2.24, 2.62]$, so we only keep the following four small intervals that have non-empty intersection with $\mathbf{E}$: $[x_{(4)}, x_{(5)}] = [2.2, 2.4]$, $[x_{(5)}, x_{(6)}] = [2.4, 2.5]$, $[x_{(6)}, x_{(7)}] = [2.5, 2.6]$, and $[x_{(7)}, x_{(8)}] = [2.6, 2.7]$. For these intervals:

- $S_4 = 7.0$, $N_4 = 3$, so $r_4 = 2.333\ldots$; $S_5 = 4.6$, $N_5 = 2$, so $r_5 = 2.3$;
- $S_6 = 2.1$, $N_6 = 1$, so $r_6 = 2.1$; $S_7 = 4.7$, $N_7 = 2$, so $r_7 = 2.35$.

Of the four values r_k, only r_4 lies within the corresponding small interval. For this small interval, $V_4' = 0.021666\ldots$, so $\underline{V} = 0.021666\ldots$

4 A Feasible Algorithm That Computes $\overline{V}$ in Many Situations

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 $\overline{V}$ for many reasonable situations. Namely, we propose an efficient algorithm $\mathcal{A}$ that computes $\overline{V}$ for the case when all the interval midpoints ("measured values") $\widetilde{x}_i = (\underline{x}_i + \overline{x}_i)/2$ are definitely different from each other, in the sense that the "narrowed" intervals $[\widetilde{x}_i - \Delta_i/n, \widetilde{x}_i + \Delta_i/n]$ – where $\Delta_i = (\underline{x}_i - \overline{x}_i)/2$ is the interval's half-width – do not intersect with each other.

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

- First, we sort all $2n$ endpoints of the narrowed intervals $\widetilde{x}_i - \Delta_i/n$ and $\widetilde{x}_i + \Delta_i/n$ 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") $[x_{(k)}, x_{(k+1)}]$, where we denote $x_{(0)} \stackrel{\text{def}}{=} -\infty$ and $x_{(2n+1)} \stackrel{\text{def}}{=} +\infty$.
- Second, we compute $\underline{E}$ and $\overline{E}$ and select all "small intervals" $[x_{(k)}, x_{(k+1)}]$ that intersect with $[\underline{E}, \overline{E}]$.
- For each of selected small intervals $[x_{(k)}, x_{(k+1)}]$, for each i from 1 to n, we pick the following value of x_i:
  - if $x_{(k+1)} < \widetilde{x}_i - \Delta_i/n$, then we pick $x_i = \overline{x}_i$;
  - if $x_{(k)} > \widetilde{x}_i + \Delta_i/n$, then we pick $x_i = \underline{x}_i$;
  - for all other i, we consider both possible values $x_i = \overline{x}_i$ and $x_i = \underline{x}_i$.

  As a result, we get one or several sequences of x_i. For each of these sequences, we check whether the average E of the selected values $x_1, \ldots, x_n$ is indeed within this small interval, and if it is, compute the sample variance by using the formula for V.
- Finally, we return the largest of the computed sample variances as $\overline{V}$.

Theorem 4.1. *The algorithm $\overline{\mathcal{A}}$ compute $\overline{V}$ in quadratic time for all the cases in which the "narrowed" intervals do not intersect with each other.*

This algorithm also works when, for some fixed k, no more than k "narrowed" intervals can have a common point:

Theorem 4.2. *For every positive integer k, the algorithm $\overline{\mathcal{A}}$ compute $\overline{V}$ is quadratic time for all the cases in which no more than k "narrowed" intervals can have a common point.*

This computation time is quadratic in n but it grows exponentially with k. So, when k 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 2.1) correspond to the case when all n narrowed intervals have a common point.

5 Sample Mean, Sample Variance: What Next?

When we have two sets of data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, we normally compute *sample covariance* $C = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$, where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. Sample covariance is used to describe the correlation between x_i and y_i. If we take interval uncertainty into consideration, then, after each measurement, we do not get the exact values of $x_1, \ldots, x_n$, $y_1, \ldots, y_n$; instead, we only have *intervals* $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]$, $[\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]$. Depending on what are the actual values of $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ within these intervals, we get different values of sample covariance. To take the interval uncertainty into consideration, we need to be able to describe the interval $[\underline{C}, \overline{C}]$ of possible values of the sample covariance C.

So, we arrive at the following problems: given the intervals $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$, compute the lower and upper bounds $\underline{C}$ and $\overline{C}$ for the interval of possible values of sample covariance.

It turns out that these problems are also NP-hard:

Theorem 5.1. *The problem of computing $\overline{C}$ from the interval inputs $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$ is NP-hard.*

Theorem 5.2. *The problem of computing $\underline{C}$ from the interval inputs $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$ is NP-hard.*

Comment. These results were first announced in [2].

As we have mentioned, sample covariance C between the data sets $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ is often used to compute sample correlation

$$\rho = \frac{C}{\sigma_x \cdot \sigma_y}, \tag{5.1}$$

where $\sigma_x = \sqrt{V_x}$ is the sample standard deviation of the values $x_1, \ldots, x_n$, and $\sigma_y = \sqrt{V_y}$ is the sample standard deviation of the values $y_1, \ldots, y_n$.

When we only have *intervals* $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]$, $[\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]$, we have an interval $[\underline{\rho}, \overline{\rho}]$ of possible value of correlation. It turns out that, similar to sample covariance, computation of the endpoints of this interval problems is also an NP-hard problem:

Theorem 5.3. *The problem of computing $\overline{\rho}$ from the interval inputs $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$ is NP-hard.*

Theorem 5.4. *The problem of computing $\underline{\rho}$ from the interval inputs $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$ is NP-hard.*

6 Proofs

Proof of Theorem 2.1

1°. To prove that our problem 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 n positive integers $s_1, \ldots, s_n$, we must check whether there exist signs $\eta_i \in \{-1, +1\}$ for which the signed sum $\sum_{i=1}^{n} \eta_i \cdot s_i$ equals 0.

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

As this instance, we take the instance corresponding to the intervals $[\underline{x}_i, \overline{x}_i] = [-s_i, s_i]$. We want to show that for the corresponding problem, $\overline{V} = C_0$, where we denoted
$C_0 \stackrel{\text{def}}{=} \frac{1}{n-1} \cdot \sum_{i=1}^{n} s_i^2$, if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

2°. Let us first show that in all cases, $\overline{V} \le C_0$.

Indeed, it is known that the formula for the sample variance can be reformulated in the following equivalent form:

$$V = \frac{1}{n-1} \cdot \sum_{i=1}^{n} x_i^2 - \frac{n}{n-1} \cdot (\bar{x})^2. \tag{6.1}$$

Since $x_i \in [-s_i, s_i]$, we can conclude that $x_i^2 \le s_i^2$ hence $\sum x_i^2 \le \sum s_i^2$. Since $(\bar{x})^2 \ge 0$, we thus conclude that $V \le \frac{1}{n-1} \cdot \sum_{i=1}^{n} s_i^2 = C_0$. In other words, every possible value V of the sample variance is smaller than or equal to C_0. Thus, the largest of these possible values, i.e., $\overline{V}$, also cannot exceed C_0, i.e., $\overline{V} \le C_0$.

3°. Let us now show that if $\overline{V} = C_0$, then the desired signs exist.

Indeed, let $\overline{V} = C_0$. Sample variance is a continuous function on a bounded closed set $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n \subset R^n$, hence its maximum on this set is attained for some values $x_1 \in \mathbf{x}_1 = [-s_1, s_1], \ldots, x_n \in \mathbf{x}_n = [-s_n, s_n]$. In other words, for the corresponding values of x_i, the sample variance V is equal to C_0.

Since $x_i \in [-s_i, s_i]$, we can conclude that $x_i^2 \le s_i^2$; since $(\bar{x})^2 \ge 0$, we get $V \le C_0$. If $x_i^2 < s_i^2$ or $(\bar{x})^2 > 0$, then we would have $V < C_0$. Thus, the only way to have $V = C_0$ is to have $x_i^2 = s_i^2$ and $\bar{x} = 0$. The first equality leads to $x_i = \pm s_i$, i.e., to $x_i = \eta_i \cdot s_i$ for some $\eta_i \in \{-1, +1\}$. The second equality then leads to $\sum_{i=1}^{n} \eta_i \cdot s_i = 0$. So, if $\overline{V} = C_0$, then the desired signs do exist.

4°. To complete the proof of Theorem 2.1, we must show that, vice versa, if the desired signs η_i exist, then $\overline{V} = C_0$.

Indeed, in this case, for $x_i = \eta_i \cdot s_i$, we have $\bar{x} = 0$ and $x_i^2 = s_i^2$, hence

$$V = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} s_i^2 = C_0.$$

The theorem is proven.

Proof of Theorem 2.2: Main Idea

The main idea of this proof is that if when the variance is close to C_0, then the values x_i are close to $\pm s_i$, and the sum $\sum x_i$ is close to 0, hence, the corresponding integers $\pm s_i$ solve the subset problem.

If ε is too large, we make the same conclusion with intervals $[-k \cdot s_i, k \cdot s_i]$ for some sufficiently large k.

Proof of Theorem 3.1

1°. Let us first show that the algorithm described in Section 3 is indeed correct.

Indeed, let $x_1^{(0)} \in \mathbf{x}_1, \ldots, x_n^{(0)} \in \mathbf{x}_n$ be the values for which the sample variance V attains minimum on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$.

If we fix the values $x_j = x_j^{(0)}$ of all the variables but one x_i, then V becomes a quadratic function of x_i. This function of one variable should attain its minimum on the interval $\mathbf{x}_i$ at the value $x_i^{(0)}$.

This function of one variable attains its *global* minimum when $\partial V / \partial x_i = 0$. Differentiating the formula for V with respect to x_i, we conclude that $\frac{\partial V}{\partial x_i} = \frac{2}{n-1} \cdot \left((x_i - E) + \sum_{j=1}^{n} (E - x_j^{(0)}) \cdot \frac{1}{n} \right) = \frac{2}{n-1} \cdot (x_i - E)$. So, this function attains the minimum when $x_i = E$.

Since

$$E = \frac{x_i}{n} + \frac{\sum_i' x_j^{(0)}}{n},$$

where $\sum_i'$ means the sum over all $j \neq i$, the equality $x_i = E$ implies that $x_i = E_i' \stackrel{\text{def}}{=} \frac{1}{n-1} \cdot \sum_i' x_j^{(0)}$, i.e., when x_i is equal to the arithmetic average E_i' of all other elements.

For $x_i < E_i'$, the function V is decreasing, for $x_i > E_i'$, this function in increasing. Thus:

- If $E_i' \in \mathbf{x}_i$, the global minimum of the function V of one variable is attained within the interval $\mathbf{x}_i$, hence the minimum on the interval $\mathbf{x}_i$ is attained for $x_i = E_i'$.
- If $E_i' < \underline{x}_i$, the function V is increasing on the interval $\mathbf{x}_i$ and therefore, its minimum on this interval is attained when $x_i = \underline{x}_i$.
- Finally, if $E_i' > \overline{x}_i$, the function V is decreasing on the interval $\mathbf{x}_i$ and therefore, its minimum on this interval is attained when $x_i = \overline{x}_i$.

Let us reformulate these conditions in terms of the average $E = \frac{1}{n} \cdot x_i + \frac{n-1}{n} \cdot E_i'$.

- In the first case, when $x_i = E_i'$, we have $x_i = E = E_i'$, so $E \in \mathbf{x}_i$.
- In the second case, we have $E_i' < \underline{x}_i$ and $x_i = \underline{x}_i$. Therefore, in this case, $E < \underline{x}_i$.
- In the third case, we have $E_i' > \overline{x}_i$ and $x_i = \overline{x}_i$. Therefore, in this case, $E > \underline{x}_i$.

Thus:

- If $E \in \mathbf{x}_i$, then we cannot be in the second or third cases. Thus, we are in the first case, hence $x_i = E$.
- If $E < \underline{x}_i$, then we cannot be in the first or the third cases. Thus, we are the second case, hence $x_i = \underline{x}_i$.
- If $E > \overline{x}_i$, then we cannot be in the first or the second cases. Thus, we are in the third case, hence $x_i = \overline{x}_i$.

Thence, as soon as we determine the position of E with respect to all the bounds $\underline{x}_i$ and $\overline{x}_i$, we will have a pretty good understanding of all the values x_i at which the minimum is attained.

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 $E \notin \mathbf{x}_i$, the values x_i that correspond to the minimal sample variance are uniquely determined by the above formulas.

For the i's for which $E \in \mathbf{x}_i$ the selected value x_i should be equal to E. To determine this E, we can use the fact that E is equal to the average of all thus selected values x_i, in other words, that we should have

$$E = \frac{1}{n} \cdot \left(\sum_{i: \underline{x}_i \geq x_{(k+1)}} \underline{x}_i + (n - N_k) \cdot E + \sum_{j: \overline{x}_j \leq x_{(k)}} \overline{x}_j \right), \tag{6.2}$$

where $(n - N_k) \cdot E$ combines all the points for which $E \in \mathbf{x}_i$. Multiplying both sides of (6.2) by n and subtracting $n \cdot E$ from both sides, we conclude that (in notations of Section 3), we have $E = S_k / N_k$ – what we denoted, in the algorithm's description, by r_k. If thus defined r_k does not

belong to the subinterval $[x_{(k)}, x_{(k+1)}]$, this contradiction with our initial assumption shows that there cannot be any minimum in this subinterval, so this subinterval can be easily dismissed.

The corresponding sample variance is denoted by V_k'. If $N_k = 0$, this means that E belongs to all the intervals $\mathbf{x}_i$ and therefore, that the lower endpoint $\underline{V}$ is exactly 0 – so we assign $V_k' = 0$.

2°. To complete the proof of Theorem 3.1, 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 Theorems 4.1 and 4.2

1°. Similarly to the proof of Theorem 3.1, let us first show that the algorithm described in Section 4 is indeed correct.

2°. Similarly to the proof of Theorem 3.1, let $x_1, \ldots, x_n$ be the values at which the sample variance attain its maximum on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$. If we fix the values of all the variables but one x_i, then V becomes a quadratic function of x_i. When the function V attains maximum over $x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n$, then this quadratic function of one variable will attain its maximum on the interval $\mathbf{x}_i$ at the point x_i.

We have already shown, in the proof of Theorem 3.1, that this quadratic function has a (global) minimum at $x_i = E_i'$, where E_i' is the average of all the values $x_1, \ldots, x_n$ except for x_i. Since this quadratic function of one variable is always non-negative, it cannot have a global maximum. Therefore, its maximum on the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ is attained at one of the endpoints of this interval.

An arbitrary quadratic function of one variable is symmetric with respect to the location of its global minimum, so its maximum on any interval is attained at the point which is the farthest from the minimum. There is exactly one point which is equally close to both endpoints of the interval $\mathbf{x}_i$: its midpoint $\widetilde{x}_i$. Depending on whether the global minimum is to the left, to the right, or exactly at the midpoint, we get the following three possible cases:

1. If the global minimum E_i' is to the left of the midpoint $\widetilde{x}_i$, i.e., if $E_i' < \widetilde{x}_i$, then the upper endpoint is the farthest from E_i'. In this case, the maximum of the quadratic function is attained at its upper endpoint, i.e., $x_i = \overline{x}_i$.

2. Similarly, if the global minimum E_i' is to the right of the midpoint $\widetilde{x}_i$, i.e., if $E_i' > \widetilde{x}_i$, then the lower endpoint is the farthest from E_i'. In this case, the maximum of the quadratic function is attained at its lower endpoint, i.e., $x_i = \underline{x}_i$.

3. If $E_i' = \widetilde{x}_i$, then the maximum of V is attained at both endpoints of the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.

3°. In the third case, we have either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$. Depending on whether x_i is equal to the lower or to the upper endpoints, we can "combine" the corresponding situations with Cases 1 and 2. As a result, we arrive at the conclusion that one of the following two situations happen:

1. either $E_i' \le \widetilde{x}_i$ and $x_i = \overline{x}_i$;

2. either $E_i' \ge \widetilde{x}_i$ and $x_i = \underline{x}_i$.

4°. Similarly to the proof of Theorem 3.1, let us reformulate these conclusions in terms of the average E of the maximizing values $x_1, \ldots, x_n$.

The average E_i' can be described as

$$\frac{\sum_i' x_j}{n-1},$$

where $\sum_i'$ means the sum over all $j \neq i$. By definition, $\sum_j' x_j = \sum_j x_j - x_i$, where $\sum_j x_j$ means the sum over all possible j. By definition of E, we have

$$E = \frac{\sum_j x_j}{n},$$

hence $\sum_j x_j = n \cdot E$. Therefore, $E_i' = \frac{n \cdot E - x_i}{n-1}$. Let us apply this formula to the above three cases.

4.1°. In the first case, we have $\widetilde{x}_i \geq E_i'$. So, in terms of E, we get the inequality $\widetilde{x}_i \geq \frac{n \cdot E - x_i}{n-1}$. Multiplying both sides of this inequality by $n-1$, and using the fact that in this case, $x_i = \overline{x}_i = \widetilde{x}_i + \Delta_i$, we conclude that $(n-1) \cdot \widetilde{x}_i \geq n \cdot E - \widetilde{x}_i - \Delta_i$. Moving all the terms but $n \cdot E$ to the left-hand side and dividing by E, we get the following inequality: $E \leq \widetilde{x}_i + \Delta_i/n$.

4.2°. In the second case, we have $\widetilde{x}_i \leq E_i'$. So, in terms of E, we get the inequality

$$\widetilde{x}_i \leq \frac{n \cdot E - x_i}{n-1}.$$

Multiplying both sides of this inequality by $n-1$, and using the fact that in this case, $x_i = \underline{x}_i = \widetilde{x}_i - \Delta_i$, we conclude that

$$(n-1) \cdot \widetilde{x}_i \leq n \cdot E - \widetilde{x}_i + \Delta_i.$$

Moving all the terms but $n \cdot E$ to the left-hand side and dividing by E, we get the following inequality:

$$E \geq \widetilde{x}_i - \frac{\Delta_i}{n}.$$

5°. Parts 4.1 and 4.2 of this proof can be summarized as follows:

- In Case 1, we have $E \leq \widetilde{x}_i + \Delta_i/n$ and $x_i = \overline{x}_i$.
- In Case 2, we have $E \geq \widetilde{x}_i - \Delta_i/n$ and $x_i = \underline{x}_i$.

Therefore:

- If $E < \widetilde{x}_i - \Delta_i/n$, this means that we cannot be in Case 2. So we must be in Case 1 and therefore, we must have $x_i = \overline{x}_i$.
- If $E > \widetilde{x}_i + \Delta_i/n$, this means that we cannot be in Case 1. So, we must be in Case 2 and therefore, we must have $x_i = \underline{x}_i$.

The only case when we do not know which endpoint for x_i we should choose is the case when E belongs to the narrowed interval $[\widetilde{x}_i - \Delta/n, \widetilde{x}_i + \Delta_i]$.

6°. Hence, 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 k narrowed intervals can have a common point, we

have no more than k undecided values x_i. Trying all possible combinations of lower and upper endpoints for these $\le k$ values requires $\le 2^k$ steps.

Thus, the overall number of steps is $O(2^k \cdot n^2)$. Since k is a constant, the overall number of steps is thus $O(n^2)$.

The theorem is proven.

Proof of Theorem 5.1

1°. Similarly to the proof of Theorem 2.1, we reduce a subset problem to the problem of computing $\overline{V}$.

Each instance of the subset problem is as follows: given n positive integers $s_1, \ldots, s_n$, to check whether there exist signs $\eta_i \in \{-1, +1\}$ for which the signed sum $\sum_{i=1}^{n} \eta_i \cdot s_i$ equals 0.

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

As this instance, we take the instance corresponding to the intervals $[\underline{x}_i, \overline{x}_i] = [\underline{y}_i, \overline{y}_i] = [-s_i, s_i]$. We want to to show that for the corresponding problem, $\overline{C} = C_0$ (where is the same as in the proof of Theorem 2.1) if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

2°. Let us first show that in all cases, $\overline{C} \le C_0$.

Indeed, it is known that the sample covariance C is bounded by the product $\sigma_x \cdot \sigma_y$ of sample standard deviations $\sigma_x = \sqrt{V_x}$ and $\sigma_y = \sqrt{V_y}$ of x and y. In the proof of Theorem 2.1, we have already proven that the sample variance V_x of the values $x_1, \ldots, x_n$ satisfies the inequality $V_x \le C_0$; similarly, the sample variance V_y of the values $y_1, \ldots, y_n$ satisfies the inequality $V_y \le C_0$. Hence, $C \le \sigma_x \cdot \sigma_y \le \sqrt{C_0} \cdot \sqrt{C_0} = C_0$. In other words, every possible value C of the sample covariance is smaller than or equal to C_0. Thus, the largest of these possible values, i.e., $\overline{C}$, also cannot exceed C_0, i.e., $\overline{C} \le C_0$.

3°. Let us now show that if $\overline{C} = C_0$, then the desired signs exist.

Indeed, if $\overline{C} = C$, this means that for the corresponding values of x_i and y_i, the sample covariance C is equal to C_0, i.e.,

$$C = C_0 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} s_i^2.$$

On the other hand, we have shown that in all cases (and in this case in particular), $C \le \sigma_x \cdot \sigma_y \le \sqrt{C_0} \cdot \sqrt{C_0} = C_0$. If $\sigma_x < \sqrt{C_0}$, then we would have $C < C_0$. So, if $C = C_0$, we have $\sigma_x = \sigma_y = \sqrt{C_0}$, i.e., $V_x = V_y = C_0$. We have already shown, in the proof of Theorem 2.1, that in this case the desired signs exist.

4°. To complete the proof of Theorem 5.1, we must show that, vice versa, if the desired signs η_i exist, then $\overline{C} = C_0$.

Indeed, in this case, for $x_i = y_i = \eta_i \cdot s_i$, we have $\bar{x} = \bar{y} = 0$ and $x_i \cdot y_i = s_i^2$, hence

$$C = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} s_i^2 = C_0.$$

The theorem is proven.

Proof of Theorem 5.2

This proof is similar to the proof of Theorem 5.1, with the only difference that in this case, we use the other part of the inequality $|C| \le \sigma_x \cdot \sigma_y$, namely, that $C \ge -\sigma_x \cdot \sigma_y$, and in the last part of the proof, we take $y_i = -x_i$.

Proof of Theorems 5.3 and 5.4: Main Idea

Similarly to the proof of Theorems 2.1 and 5.1, we reduce a subset problem to the problem of computing $\overline{V}$. Namely, we reduce each instance of the subset problem to the following instance of our problem:

- $n = m + 2$ (note the difference between this reduction and reductions from the proofs of Theorems 2.1 and 5.1, where we have $n = m$);
- $[\underline{x}_i, \overline{x}_i] = [-s_i, s_i]$ and $\mathbf{y}_i = [0, 0]$ for $i = 1, \ldots, m$; $\mathbf{x}_{m+1} = \mathbf{y}_{m+2} = [1, 1]$; $\mathbf{x}_{m+2} = \mathbf{y}_{m+1} = [-1, -1]$.

Like in the proof of Theorem 2.1, we define C_1 as $C_1 = \sum_{i=1}^{m} s_i^2$. We can then prove that for the corresponding problem, $\overline{\rho} = -\sqrt{\frac{2}{C_1 + 2}}$ if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$. Indeed, one can show that in this case, $\rho = -\sqrt{\frac{2}{(m+1) \cdot V_x}}$. Therefore, $\rho \to \max$ iff V_x takes the largest possible value $\overline{V}_x$.

This proof is similar to the proof of Theorem 5.3, with the only difference that we take $y_{m+1} = 1$ and $y_{m+2} = -1$. In this case, $\rho = \sqrt{\frac{2}{(m+1) \cdot V_x}}$.

Acknowledgments. This work was supported in part by NASA grants NCC5-209 and NCC 2-1232, by NSF grants CDA-9522207, ERA-0112968, and 9710940 Mexico/Conacyt, by the AFOSR grant F49620-00-1-0365, and by the NIH grant 9R44CA81741.

References

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

[2] 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 (to appear).

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

[4] G. W. Walster, "Philosophy and practicalities of interval arithmetic", In: R. E. Moore (ed.), *Reliability in Computing*, 1988, pp. 307–323.