

# Population Variance under Interval Uncertainty: A New Algorithm

Evgeny Dantsin<sup>1</sup>, Vladik Kreinovich<sup>2</sup>,  
Alexander Wolpert<sup>1</sup>, and Gang Xiang<sup>2</sup>

<sup>1</sup>Department of Computer Science, Roosevelt University  
Chicago, IL 60605, USA, {edantsin,awolpert}@roosevelt.edu

<sup>2</sup>Department of Computer Science, University of Texas at El Paso,  
El Paso, TX 79968, USA, {vladik,gxiang}@utep.edu

## Abstract

In statistical analysis of measurement results, it is often beneficial to compute the range  $\mathbf{V}$  of the population variance  $V = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - E)^2$

(where  $E = \frac{1}{n} \sum_{i=1}^n x_i$ ) when we only know the intervals

$$[\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$$

of possible values of the  $x_i$ . In general, this problem is NP-hard; a polynomial-time algorithm is known for the case when the measurements are sufficiently accurate, i.e., when  $|\tilde{x}_i - \tilde{x}_j| \geq \frac{\Delta_i}{n} + \frac{\Delta_j}{n}$  for all  $i \neq j$ . In this paper, we show that we can efficiently compute  $\mathbf{V}$  under a weaker (and more general) condition  $|\tilde{x}_i - \tilde{x}_j| \geq \frac{|\Delta_i - \Delta_j|}{n}$ .

**Formulation of the problem.** Once we have  $n$  measurement results  $x_1, \dots, x_n$ , the traditional statistical analysis starts with computing the standard statistics such as population mean  $E = \frac{1}{n} \cdot \sum_{i=1}^n x_i$  and population variance

$$V = M - E^2, \text{ where } M \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^n x_i^2; \text{ see, e.g., [7].}$$

In many real-life situations, due to measurement uncertainty, instead of the actual values  $x_i$  of the measured quantity, we only have intervals  $\mathbf{x}_i = [\underline{x}_i, \bar{x}_i]$  of possible values of  $x_i$  [5, 7]. Usually, the interval  $\mathbf{x}_i$  has the form  $[\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$ ,

where  $\tilde{x}_i$  is the measurement result, and  $\Delta_i$  is the known upper bound on the absolute value of the measurement error  $\Delta x_i \stackrel{\text{def}}{=} \tilde{x}_i - x_i$ .

Different values  $x_i \in \mathbf{x}_i$  lead, in general, to different values of  $E$  and  $V$ . It is therefore desirable to compute the ranges  $\mathbf{E} = [\underline{E}, \overline{E}]$  and  $\mathbf{V} = [\underline{V}, \overline{V}]$  of possible values of  $E$  and  $V$  when  $x_i \in \mathbf{x}_i$ .

Since the population mean  $E$  is a monotonic function of its  $n$  variables  $x_1, \dots, x_n$ , its range can be easily computed as  $\mathbf{E} = \left[ \frac{1}{n} \cdot \sum_{i=1}^n x_i, \frac{1}{n} \cdot \sum_{i=1}^n \tilde{x}_i \right]$ . For the variance  $V$ , there exist polynomial-time algorithms for computing the lower bound  $\underline{V}$ , but computing the exact upper bound  $\overline{V}$  is, in general, an NP-hard problem; see, e.g., [2, 3].

There exist polynomial-time algorithms for computing  $\overline{V}$  in many practically reasonable situations; see, e.g., [2, 3, 4, 6, 8]. One such known case is when measurements are sufficiently accurate, e.g., when the “narrowed intervals”

$$\left[ \tilde{x}_i - \frac{\Delta_i}{n}, \tilde{x}_i + \frac{\Delta_i}{n} \right] \quad (1)$$

do not intersect. In other words, we know how to efficiently compute  $\overline{V}$  when for every  $i \neq j$ , we have

$$|\tilde{x}_i - \tilde{x}_j| \geq \frac{\Delta_i}{n} + \frac{\Delta_j}{n}. \quad (2)$$

The known algorithm requires  $O(n \cdot \log(n))$  computational steps.

In this paper, we propose a new algorithm that computes  $\overline{V}$  in  $O(n \cdot \log(n))$  time under the weaker (hence more general) condition

$$|\tilde{x}_i - \tilde{x}_j| \geq \frac{|\Delta_i - \Delta_j|}{n}. \quad (3)$$

This condition is indeed much weaker: e.g., for the case when all measurements are equally accurate, i.e.,  $\Delta_i = \Delta$  for all  $i$ , the previously known condition (2) is only valid for  $\Delta \leq (n/2) \cdot \min_{i \neq j} |\tilde{x}_i - \tilde{x}_j|$ , while the new condition (3) holds for every  $\Delta$ . Thus, we can have larger measurement uncertainty  $\Delta$  than before and still be able to compute the exact bound  $\overline{V}$  in polynomial time.

**Algorithm.** Let us first describe the algorithm itself; in the next section, we provide the justification for this algorithm.

- First, we sort of the values  $\tilde{x}_i$  into an increasing sequence. Without losing generality, we can assume that  $\tilde{x}_1 \leq \tilde{x}_2 \leq \dots \leq \tilde{x}_n$ .
- Then, for every  $k$  from 0 to  $n$ , we compute the value  $V^{(k)} = M^{(k)} - E^{(k)}$  of the population variance  $V$  for the vector  $x^{(k)} = (\underline{x}_1, \dots, \underline{x}_k, \overline{x}_{k+1}, \dots, \overline{x}_n)$ .
- Finally, we compute  $\overline{V}$  as the largest of  $n + 1$  values  $V^{(0)}, \dots, V^{(n)}$ .

To compute the values  $V^{(k)}$ , first, we explicitly compute  $M^{(0)}$ ,  $E^{(0)}$ , and  $V^{(0)} = M^{(0)} - (E^{(0)})^2$ . Once we know the values  $M^{(k)}$  and  $E^{(k)}$ , we can compute  $M^{(k+1)} = M^{(k)} + \frac{1}{n} \cdot (\underline{x}_{k+1})^2 - \frac{1}{n} \cdot (\bar{x}_{k+1})^2$  and  $E^{(k+1)} = E^{(k)} + \frac{1}{n} \cdot \underline{x}_{k+1} - \frac{1}{n} \cdot \bar{x}_{k+1}$ .

**Number of computation steps.** Sorting requires  $O(n \cdot \log(n))$  steps; see, e.g., [1]. Computing the initial values  $M^{(0)}$ ,  $E^{(0)}$ , and  $V^{(0)}$  requires linear time  $O(n)$ . For each  $k$  from 0 to  $n-1$ , we need a constant number of steps to compute the next values  $M^{(k+1)}$ ,  $E^{(k+1)}$ , and  $V^{(k+1)}$ . Finally, finding the largest of  $n+1$  values  $V^{(k)}$  also requires  $O(n)$  steps. Thus, overall, we need

$$O(n \cdot \log(n)) + O(n) + O(n) + O(n) = O(n \cdot \log(n))$$

steps.

It is worth mentioning that if the measurement results  $\tilde{x}_i$  are already sorted, then we only need linear time to compute  $\bar{V}$ .

**Justification of the algorithm.** With respect to each variable  $x_i$ , the population variance is a quadratic function which is non-negative for all  $x_i$ . It is well known that a maximum of such a function on each interval  $[\underline{x}_i, \bar{x}_i]$  is attained at one of the endpoints of this interval. Thus, the maximum  $\bar{V}$  of the population variance is attained at a vector  $x = (x_1, \dots, x_n)$  in which each value  $x_i$  is equal either to  $\underline{x}_i$  or to  $\bar{x}_i$ .

We will first justify our algorithm for the case when  $|\tilde{x}_i - \tilde{x}_j| > \frac{|\Delta_i - \Delta_j|}{n}$  for all  $i \neq j$ .

To justify our algorithm, we need to prove that this maximum is attained at one of the vectors  $x^{(k)}$  in which all the lower bounds  $\underline{x}_i$  precede all the upper bounds  $\bar{x}_i$ . We will prove this by reduction to a contradiction. Indeed, let us assume that the maximum is attained at a vector  $x$  in which one of the lower bounds follows one of the upper bounds. In each such vector, let  $i$  be the largest upper bound index preceded by the lower bound; then, in the optimal vector  $x$ , we have  $x_i = \bar{x}_i$  and  $x_{i+1} = \underline{x}_{i+1}$ .

Since the maximum is attained for  $x_i = \bar{x}_i$ , replacing it with  $\underline{x}_i = \bar{x}_i - 2 \cdot \Delta_i$  will either decrease the value of the variance or keep it unchanged. Let us describe how variance changes under this replacement. In the sum for  $M$ , we replace  $(\bar{x}_i)^2$  with

$$(\underline{x}_i)^2 = (\bar{x}_i - 2 \cdot \Delta_i)^2 = (\bar{x}_i)^2 - 4 \cdot \Delta_i \cdot \bar{x}_i + 4 \cdot \Delta_i^2.$$

Thus, the value  $M$  changes into  $M + \Delta M_i$ , where

$$\Delta M_i = -\frac{4}{n} \cdot \Delta_i \cdot \bar{x}_i + \frac{4}{n} \cdot \Delta_i^2.$$

The population mean  $E$  changes into  $E + \Delta E_i$ , where  $\Delta E_i = -\frac{2 \cdot \Delta_i}{n}$ . Thus, the value  $E^2$  changes into  $(E + \Delta E_i)^2 = E^2 + \Delta(E^2)_i$ , where

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

So, the variance  $V$  changes into  $V + \Delta V_i$ , where

$$\begin{aligned} \Delta V_i &= \Delta M_i - \Delta(E^2)_i = -\frac{4}{n} \cdot \Delta_i \cdot \bar{x}_i + \frac{4}{n} \cdot \Delta_i^2 + \frac{4}{n} \cdot E \cdot \Delta_i - \frac{4}{n^2} \cdot \Delta_i^2 = \\ &= \frac{4}{n} \cdot \Delta_i \cdot \left( -\bar{x}_i + \Delta_i + E - \frac{\Delta_i}{n} \right). \end{aligned}$$

By definition,  $\bar{x}_i = \tilde{x}_i + \Delta_i$ , hence  $-\bar{x}_i + \Delta_i = -\tilde{x}_i$ . Thus, we conclude that

$$\Delta V_i = \frac{4}{n} \cdot \Delta_i \cdot \left( -\tilde{x}_i + E - \frac{\Delta_i}{n} \right).$$

Since  $V$  attains maximum at  $x$ , we have  $\Delta V_i \leq 0$ , hence

$$E \leq \tilde{x}_i + \frac{\Delta_i}{n}. \quad (4)$$

Similarly, since the maximum is attained for  $x_{i+1} = \underline{x}_i$ , replacing it with  $\bar{x}_{i+1} = \underline{x}_{i+1} + 2 \cdot \Delta_{i+1}$  will either decrease the value of the variance or keep it unchanged. Let us describe how variance changes under this replacement. In the sum for  $M$ , we replace  $(\underline{x}_{i+1})^2$  with

$$(\bar{x}_{i+1})^2 = (\underline{x}_{i+1} + 2 \cdot \Delta_{i+1})^2 = (\underline{x}_{i+1})^2 + 4 \cdot \Delta_{i+1} \cdot \underline{x}_{i+1} + 4 \cdot \Delta_{i+1}^2.$$

Thus, the value  $M$  changes into  $M + \Delta M_{i+1}$ , where

$$\Delta M_{i+1} = \frac{4}{n} \cdot \Delta_{i+1} \cdot \underline{x}_{i+1} + \frac{4}{n} \cdot \Delta_{i+1}^2.$$

The population mean  $E$  changes into  $E + \Delta E_{i+1}$ , where  $\Delta E_{i+1} = \frac{2 \cdot \Delta_{i+1}}{n}$ . Thus, the value  $E^2$  changes into  $(E + \Delta E_{i+1})^2 = E^2 + \Delta(E^2)_{i+1}$ , where

$$\Delta(E^2)_{i+1} = 2 \cdot E \cdot \Delta E_{i+1} + \Delta E_{i+1}^2 = \frac{4}{n} \cdot E \cdot \Delta_{i+1} + \frac{4}{n^2} \cdot \Delta_{i+1}^2.$$

So, the variance  $V$  changes into  $V + \Delta V_{i+1}$ , where

$$\begin{aligned} \Delta V_{i+1} &= \Delta M_{i+1} - \Delta(E^2)_{i+1} = \\ &= \frac{4}{n} \cdot \Delta_{i+1} \cdot \underline{x}_{i+1} + \frac{4}{n} \cdot \Delta_{i+1}^2 - \frac{4}{n} \cdot E \cdot \Delta_{i+1} - \frac{4}{n^2} \cdot \Delta_{i+1}^2 = \end{aligned}$$

$$\frac{4}{n} \cdot \Delta_{i+1} \cdot \left( \underline{x}_{i+1} + \Delta_{i+1} - E - \frac{\Delta_{i+1}}{n} \right).$$

By definition,  $\underline{x}_{i+1} = \tilde{x}_{i+1} - \Delta_{i+1}$ , hence  $\underline{x}_{i+1} + \Delta_{i+1} = \tilde{x}_{i+1}$ . Thus, we conclude that

$$\Delta V_{i+1} = \frac{4}{n} \cdot \Delta_{i+1} \cdot \left( \tilde{x}_{i+1} - E - \frac{\Delta_{i+1}}{n} \right).$$

Since  $V$  attains maximum at  $x$ , we have  $\Delta V_{i+1} \leq 0$ , hence

$$E \geq \tilde{x}_{i+1} - \frac{\Delta_{i+1}}{n}. \quad (5)$$

We can also change *both*  $x_i$  and  $x_{i+1}$  at the same time. In this case, the change  $\Delta M$  in  $M$  is simply the sum of the changes coming from  $x_i$  and  $x_{i+1}$ :  $\Delta M = \Delta M_i + \Delta M_{i+1}$ , and the change  $\Delta E$  in  $E$  is also the sum of the corresponding changes:  $\Delta E = \Delta E_i + \Delta E_{i+1}$ . So, for

$$\Delta V = \Delta M - \Delta(E^2) = \Delta M - 2 \cdot E \cdot \Delta E - \Delta E^2,$$

we get

$$\begin{aligned} \Delta V &= \Delta M_i + \Delta M_{i+1} - \\ &2 \cdot E \cdot \Delta E_i - 2 \cdot E \cdot \Delta E_{i+1} - (\Delta E_i)^2 - (\Delta E_{i+1})^2 - 2 \cdot \Delta E_i \cdot \Delta E_{i+1}. \end{aligned}$$

Hence,

$$\begin{aligned} \Delta V &= (\Delta M_i - 2 \cdot E \cdot \Delta E_i - (\Delta E_i)^2) + (\Delta M_{i+1} - 2 \cdot E \cdot \Delta E_{i+1} - (\Delta E_{i+1})^2) \\ &\quad - 2 \cdot \Delta E_i \cdot \Delta E_{i+1}, \end{aligned}$$

i.e.,

$$\Delta V = \Delta V_i + \Delta V_{i+1} - 2 \cdot \Delta E_i \cdot \Delta E_{i+1}.$$

We already have the expressions for  $\Delta V_i$ ,  $\Delta V_{i+1}$ ,  $\Delta E_i = -\frac{2 \cdot \Delta_i}{n}$ , and  $\Delta E_{i+1} = \frac{2 \cdot \Delta_{i+1}}{n}$ , so we conclude that  $\Delta V = \frac{4}{n} \cdot D(E)$ , where

$$D(E) \stackrel{\text{def}}{=} \Delta_i \cdot \left( -\tilde{x}_i + E - \frac{\Delta_i}{n} \right) + \Delta_{i+1} \cdot \left( \tilde{x}_{i+1} - E - \frac{\Delta_{i+1}}{n} \right) + \frac{2}{n} \cdot \Delta_i \cdot \Delta_{i+1}. \quad (6)$$

Since the function  $V$  attains maximum at  $x$ , we have  $\Delta V \leq 0$ , hence  $D(E) \leq 0$  (for the population mean  $E$  corresponding to the optimizing vector  $x$ ).

The expression  $D(E)$  is a linear function of  $E$ . From (4) and (5), we know that

$$\tilde{x}_{i+1} - \frac{\Delta_{i+1}}{n} \leq E \leq \tilde{x}_i + \frac{\Delta_i}{n}.$$

For  $E = E^- \stackrel{\text{def}}{=} \tilde{x}_{i+1} - \frac{\Delta_{i+1}}{n}$ , we have

$$\begin{aligned} D(E^-) &= \Delta_i \cdot \left( -\tilde{x}_i + \tilde{x}_{i+1} - \frac{\Delta_{i+1}}{n} - \frac{\Delta_i}{n} \right) + \frac{2}{n} \cdot \Delta_i \cdot \Delta_{i+1} = \\ &\quad \Delta_i \cdot \left( -\tilde{x}_i + \tilde{x}_{i+1} + \frac{\Delta_{i+1}}{n} - \frac{\Delta_i}{n} \right). \end{aligned}$$

We consider the case when  $|\tilde{x}_{i+1} - x_i| > \frac{|\Delta_i - \Delta_{i+1}|}{n}$ . Since the values  $\tilde{x}_i$  are sorted in increasing order, we have  $\tilde{x}_{i+1} \geq \tilde{x}_i$ , hence

$$\tilde{x}_{i+1} - \tilde{x}_i = |\tilde{x}_{i+1} - \tilde{x}_i| > \frac{|\Delta_i - \Delta_{i+1}|}{n} \geq \frac{\Delta_i}{n} - \frac{\Delta_{i+1}}{n}.$$

So, we conclude that  $D(E^-) > 0$ .

For  $E = E^+ \stackrel{\text{def}}{=} \tilde{x}_i + \frac{\Delta_i}{n}$ , we have

$$\begin{aligned} D(E^+) &= \Delta_{i+1} \cdot \left( \tilde{x}_{i+1} - \tilde{x}_i - \frac{\Delta_i}{n} - \frac{\Delta_{i+1}}{n} \right) + \frac{2}{n} \cdot \Delta_i \cdot \Delta_{i+1} = \\ &\quad \Delta_{i+1} \cdot \left( -\tilde{x}_i + \tilde{x}_{i+1} + \frac{\Delta_i}{n} - \frac{\Delta_{i+1}}{n} \right). \end{aligned}$$

Here, from  $|\tilde{x}_{i+1} - x_i| > \frac{|\Delta_i - \Delta_{i+1}|}{n}$ , we also conclude that  $D(E^+) > 0$ .

Since the linear function  $D(E)$  is positive on both endpoints of the interval  $[E^-, E^+]$ , it must be positive for every value  $E$  from this interval, which contradicts to our conclusion that  $D(E) \geq 0$  for the actual population mean value  $E \in [E^-, E^+]$ . This contradiction shows that the maximum of the population variance  $V$  is indeed attained at one of the values  $x^{(k)}$ , hence the algorithm is justified.

The general case when  $|\tilde{x}_i - \tilde{x}_j| \geq \frac{|\Delta_i - \Delta_j|}{n}$  can be obtained as a limit of cases when we have strict inequality. Since the function  $V$  is continuous, the value  $\bar{V}$  continuously depends on the input bounds, so by tending to a limit, we can conclude that our algorithm works in the general case as well.

**The geometric meaning of the new condition.** The condition  $|\tilde{x}_i - \tilde{x}_j| \geq \frac{|\Delta_i - \Delta_j|}{n}$  means that if  $\tilde{x}_i \geq \tilde{x}_j$ , then we have

$$\tilde{x}_i - \tilde{x}_j \geq \frac{\Delta_i - \Delta_j}{n},$$

i.e.,

$$\tilde{x}_i - \frac{\Delta_i}{n} \geq \tilde{x}_j - \frac{\Delta_j}{n}$$

and also

$$\tilde{x}_i - \tilde{x}_j \geq \frac{\Delta_j - \Delta_i}{n},$$

i.e.,

$$\tilde{x}_i + \frac{\Delta_i}{n} \geq \tilde{x}_j + \frac{\Delta_j}{n}.$$

This means that no narrowed interval (1) is a proper subinterval of the interior of another narrowed subinterval.

Vice versa, if one of the narrowed intervals is a proper subinterval of another one, then the condition (3) is not satisfied. Thus, the condition (3) means that no *narrowed subintervals* are proper subintervals of each other.

It is worth mentioning that there is another polynomial-time algorithm for computing  $\bar{V}$  [6] – an algorithm which computes  $\bar{V}$  for the case when no *intervals* are proper subintervals of each other. That condition can be similarly described as  $|\tilde{x}_i - \tilde{x}_j| \geq |\Delta_i - \Delta_j|$ , hence that condition implies our condition (2). So, our algorithm generalizes that algorithm as well.

**Acknowledgments.** This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grant EAR-0225670, NIH grant 3T34GM008048-20S1, and Army Research Lab grant DATM-05-02-C-0046. This work was mainly done during E. Dantsin’s and A. Wolpert’s visit to El Paso. The authors are thankful to Luc Longpré for valuable discussions.

## References

- [1] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2001.
- [2] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Computing Variance for Interval Data is NP-Hard, *ACM SIGACT News*, 2002, Vol. 33, No. 2, pp. 108–118.
- [3] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Exact Bounds on Finite Populations of Interval Data, *Reliable Computing*, 2005, Vol. 11, No. 3, pp. 207–233.
- [4] L. Granvilliers, V. Kreinovich, and N. Müller, Novel Approaches to Numerical Software with Result Verification, In: R. Alt, A. Frommer, R. B. Kearfott, and W. Luther, editors, *Numerical Software with Result Verification*, Proceedings of the International Dagstuhl Seminar, Dagstuhl Castle, Germany, January 19–24, 2003, Springer Lectures Notes in Computer Science, 2004, Vol. 2991, pp. 274–305.

- [5] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied interval analysis: with examples in parameter and state estimation, robust control and robotics*, Springer Verlag, London, 2001.
- [6] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos, “Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity”, *Reliable Computing* (to appear).
- [7] S. Rabinovich, *Measurement Errors: Theory and Practice*, American Institute of Physics, New York, 1993.
- [8] G. Xiang, “Fast algorithm for computing the upper endpoint of sample variance for interval data: case of sufficiently accurate measurements”, *Reliable Computing* (to appear).