Exact Bounds
on Sample Variance
of Interval Data

Scott Ferson and Lev Ginzburg
Applied Biomathematics
100 North Country Road,
Setauket, NY 11733, USA
{scott,lev}@ramas.com

Vladik Kreinovich, Luc Longpré,
and Monica Aviles
Computer Science Department
University of Texas at El Paso
El Paso, TX 79968, USA
{maviles,longpre,vladik}@cs.utep.edu
Formulation of the Problem

- We have \( n \) measurement results \( x_1, \ldots, x_n \).

- Traditional statistical approach: compute

\[
E = \bar{x} = \frac{x_1 + \ldots + x_n}{n},
\]

\[
V = \frac{(x_1 - E)^2 + \ldots + (x_n - E)^2}{n - 1} \quad \text{(or } \sigma = \sqrt{V}).
\]

- *Reasons:* \( V \) is an unbiased estimator of the variance; for Gaussian, it is MLM.

- Often, we only have intervals \( x_i = [\underline{x}_i, \bar{x}_i] \).

- *Example:* for measurements, \( x_i = [\underline{x}_i - \Delta_i, \bar{x}_i + \Delta_i] \).

- What are \( \mathbf{E} \) and \( \mathbf{V} = [\underline{V}, \bar{V}] \)?

- For \( \mathbf{E} \), the answer is easy.

- When \( \cap_{i=1}^n x_i \neq \emptyset \), we have \( V = 0 \); else \( V > 0 \).

- *Problem* (Walster): what is the total set \( \mathbf{V} \) of possible values of \( V \)?
For this Problem,
Straightforward
Interval Computations
Sometimes Overestimate

• **Reminder:**
  
  – parse the function $f(x_1, \ldots, x_n)$, and
  
  – replace each elementary operation by the corr. operation of interval arithmetic.

• **Example:** for $x_1 = x_2 = [0, 1]$.

• **Actual range:** since $V = (x_1 - x_2)^2/2$, the actual range is $V = [0, 0.5]$.

• **Estimate:** $E = [0, 1]$, hence

  $$(x_1 - E)^2 + (x_2 - E)^2 = [0, 2] \supset [0, 0.5].$$
Centered Form
Sometimes Overestimates

- **Reminder:**

\[
f(x_1, \ldots, x_n) \subseteq f(\bar{x}_1, \ldots, \bar{x}_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) \cdot [-\Delta_i, \Delta_i],
\]

where:

- \(\bar{x}_i = (\underline{x}_i + \bar{x}_i)/2\) is the interval’s midpoint and
- \(\Delta_i = (\underline{x}_i - \bar{x}_i)/2\) is its half-width.

- **Not perfect** (similar to Hertling):
  - it produces an interval centered at \(f(\bar{x}_1, \ldots, \bar{x}_n)\);
  - when all intervals \(x_i\) are equal, all midpoints \(\bar{x}_i\) are the same;
  - hence the sample variance \(f(\bar{x}_1, \ldots, \bar{x}_n)\) is 0;
  - so, the estimate’s lower bound is < 0, but \(V \geq 0\).
First Result: Computing $\mathbf{V}$

The following algorithm always compute $\mathbf{V}$ in $O(n^2)$:

- First, we sort all $2n$ values $x_i, \overline{x}_i$ into a sequence $x(1) \leq x(2) \leq \ldots \leq x(2n)$.

- Second, we compute $E$ and $\overline{E}$ and select all “small intervals” $[x(k), x(k+1)]$ that intersect with $[E, \overline{E}]$.

- For each of the selected small intervals $[x(k), x(k+1)]$, we compute the ratio $r_k = S_k/N_k$, where

$$S_k \overset{\text{def}}{=} \sum_{i : x_i \geq x(k+1)} x_i + \sum_{j : \overline{x}_j \leq x(k)} \overline{x}_j,$$

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

$$V'_k \overset{\text{def}}{=} \frac{1}{n - 1} \left( \sum_{i : x_i > x(k+1)} (x_i - r)^2 + \sum_{j : \overline{x}_j < x(k)} (\overline{x}_j - r)^2 \right).$$

If $N_k = 0$, we take $V'_k \overset{\text{def}}{=} 0$.

- Finally, we return the smallest of the values $V'_k$ as $\mathbf{V}$. 
Example

• Input: \( \mathbf{x}_1 = [2.1, 2.6], \mathbf{x}_2 = [2.0, 2.1], \mathbf{x}_3 = [2.2, 2.9], \mathbf{x}_4 = [2.5, 2.7], \) and \( \mathbf{x}_5 = [2.4, 2.8]. \)

• “small intervals”: \([x_1(1), x_1(2)] = [2.0, 2.1], [2.1, 2.1], [2.1, 2.2], [2.2, 2.4], [2.4, 2.5], [2.5, 2.6], [2.6, 2.7], [2.7, 2.8], \) and \([2.8, 2.9]. \)

• Sample average \( \mathbf{E} = [2.24, 2.62], \) so we keep \([2.2, 2.4], [2.4, 2.5], [2.5, 2.6], [2.6, 2.7]. \) 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 \( r_7 = 2.35. \)

• Only \( r_4 \) lies within the corresponding small interval.

• Here, \( V'_4 = 0.021666\ldots, \) so \( V = 0.021666\ldots \)
Second Result:
Computing $\nabla$ is NP-Hard

- **Theorem.** Computing $\nabla$ is NP-hard.

- **Comments:**
  
  - NP-hard means, crudely speaking, that there are no general ways for solving all particular cases of this problem in reasonable time.
  
  - NP-hardness of computing the range of a quadratic function was proven by Vavasis (1991).
  
  - By using peeling, we can compute $\nabla$ in exponential time $O(2^n)$.

- **Natural question:** maybe the difficulty comes from the requirement that the range be computed exactly?

- **Theorem.** For every $\varepsilon > 0$, the problem of computing $\nabla$ with accuracy $\varepsilon$ is NP-hard.
Third Result:
A Feasible Algorithm
that Computes $\mathcal{V}$
in Many Practical Situations

• **Case:** all midpoints ("measured values")

$$\bar{x}_i = \frac{x_i + \bar{x}_i}{2}$$

of the intervals

$$x_i = [\bar{x}_i - \Delta_i, \bar{x}_i + \Delta_i]$$

are definitely different from each other.

• **Namely:** the "narrowed" intervals

$$\left[\bar{x}_i - \frac{\Delta_i}{n}, \bar{x}_i + \frac{\Delta_i}{n}\right]$$

do not intersect with each other.

• In this case, there exists an algorithm computes $\mathcal{V}$ in quadratic time.
Algorithm

- Sort $2n$ endpoints of narrowed intervals into $x(1) \leq x(2) \leq \ldots \leq x(2n)$.
- Thus, $IR$ is divided into $2n + 2$ segments ("small intervals") $[x(k), x(k+1)]$.
- Select only "small intervals" $[x(k), x(k+1)]$ that intersect with $E$; for each, pick $x_i$ as follows:
  - if $x(k+1) < \bar{x}_i - \Delta_i/n$, then we pick $x_i = \bar{x}_i$;
  - if $x(k) > \bar{x}_i + \Delta_i/n$, then we pick $x_i = \bar{x}_i$;
  - for all other $i$, we consider both possible values $x_i = \bar{x}_i$ and $x_i = \bar{x}_i$.
- For each of the sequences $x_i$, we check whether the average $E$ is indeed within this small interval, and if it is, compute the sample variance.
- The largest of the computed sample variances is $\nabla$. 
Third Result (cont-d)

- **Question:** what if two “narrowed” intervals have a common point?

- **Case:** let us fix $k$ and consider all cases $C_k$ in which no more than $k$ “narrowed” intervals can have a common point.

- **Result:** $\forall k$, the above algorithm $\mathcal{A}$ computes $\mathcal{V}$ in quadratic time for all problems $\in C_k$.

- **Comments:**
  
  - Computation time $t$ is quadratic in $n$.
  
  - However, $t$ is exponential in $k$.
  
  - So, when $k \uparrow$, the algorithm $\mathcal{A}$ requires more and more computation time.
  
  - In our proof of NP-hardness, we use the case when all $n$ narrowed intervals have a common point.
Sample Mean, Sample Variance: What Next?

- **Sample covariance**
  \[
  C = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}).
  \]
- **Result:** both computing \( C \) and computing \( C \) are NP-hard problems.

- **Sample correlation**
  \[
  \rho = \frac{C}{\sigma_x \cdot \sigma_y}.
  \]
- **Result:** both computing \( \rho \) and computing \( \rho \) are NP-hard problems.

- **Open problem:** design feasible algorithms that work in many practical cases.

- **Median:** feasible (since it is monotonic in \( x_i \)).

- **Open problem:** analyze other statistical characteristics from this viewpoint.
Acknowledgments

This work was supported in part:

• by NASA under grants NCC5-209, NCC2-1232, and NCC2-1243;

• by the Air Force Office of Scientific Research grants F30602-00-2-0503 and F49620-00-1-0365;

• by grant No. W-00016 from the U.S.-Czech Science and Technology Joint Fund, and

• by NSF grants CDA-9522207, ERA-0112968 and 9710940 Mexico/Conacyt.