Estimating Mean under Interval Uncertainty and Variance Constraint

Ali Jalal-Kamali[1], Luc Longpré[1], and Misha Koshelev[2]

[1]Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
ajalalkamali@miners.utep.edu
longpre@utep.edu

[2]Human Neuroimaging Lab
Division of Neuroscience
Baylor College of Medicine
Houston, TX 77030, USA
misha680hnl@gmail.com

1. Analyzing a Sample

- Often, we have a sample of values $x_1, \ldots, x_n$ corresponding to objects of a certain type.
- In this case, a standard way to describe the corresponding population is to estimate its mean and variance:

$$E = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i; \quad V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2.$$

- In practice, the values x_i come from measurements, and measurements are never absolutely accurate.
- Often, the only information we have is an upper bound Δ_i on the measurement error: $|\Delta x_i| \le \Delta_i$.
- In this case, based on the measured value $\widetilde{x}_i$, we conclude that the actual value x_i is in the interval

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

2. Need to Estimate Mean and Variance under Interval Uncertainty

- In general, different values $x_i \in \mathbf{x}_i$ lead to different values of E and V.
- It is therefore desirable to describe the range of possible values of mean and variance when $x_i \in \mathbf{x}_i$.
- This is a particular case of a general problem of *interval computation*: computing the range

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

- Sometimes, we have fuzzy values $X_1, \ldots, X_n$, and we want to find $Y = f(X_1, \ldots, X_n)$.
- It is known that for α-cuts $X_i(\alpha)$, we have

$$Y(\alpha) = \{f(x_1, \ldots, x_n) \,|\, x_1 \in X_1(\alpha), \ldots, x_n \in X_n(\alpha)\}.$$

- In view of this reduction, we will concentrate on algorithms for interval uncertainty.

3. Computing the Ranges of the Mean and Variance: What Is Known

- The mean E is an increasing function of each x_i; thus:
  - the smallest value $\underline{E}$ is attained when each x_i is the smallest $x_i = \underline{x}_i$, and
  - the largest value $\overline{E}$ is attained when each x_i is the largest $x_i = \overline{x}_i$:

$$\underline{E} = \frac{1}{n} \cdot \sum_{i=1}^{n} \underline{x}_i; \quad \overline{E} = \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i.$$

- Variance V is, in general, not monotonic, so its range is more difficult to compute:
  - the lower endpoint $\underline{V}$ is computable in linear time,
  - but computing $\overline{V}$ is, in general, NP-hard.
- There are also efficient algorithms for computing $\overline{V}$ in some cases.

4. Variance Constraints

- In the previous expressions, we assume that there is no *a priori* information about the values of E and V.
- In some cases, we have *a priori* constraint on the variance: $V \le V_0$, for a given V_0.
- For example, we know that within a species, there can be ≤ 0.1 variation of a certain characteristic.
- Thus, we arrive at the following problem:
  - *given:* n intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ and a number $V_0 \ge 0$;
  - *compute:* the range

$$[\underline{E}, \overline{E}] = \{E(x_1, \ldots, x_n) \,:\, x_i \in \mathbf{x}_i \,\&\, V(x_1, \ldots, x_n) \le V_0\};$$

  - *under the assumption* that there exist values $x_i \in \mathbf{x}_i$ for which $V(x_1, \ldots, x_n) \le V_0$.
- This is the problem that we will solve in this paper.

5. Cases Where This Problem Is (Relatively) Easy to Solve

- *First case:* V_0 is $\geq$ the largest possible value $\overline{V}$ of the variance corresponding to the given sample.
- In this case, the constraint $V \leq V_0$ is always satisfied.
- Thus, in this case, the desired range simply coincides with the range of all possible values of E.
- *Second case:* $V_0 = 0$.
- In this case, the constraint $V \leq V_0$ means that the variance V should be equal to 0, i.e., $x_1 = \ldots = x_n$.
- In this case, we know that this common value x_i belongs to each of n intervals $\mathbf{x}_i$.
- So, the set of all possible values E is the intersection:

$$E = \mathbf{x}_1 \cap \ldots \cap \mathbf{x}_n.$$

6. Main Result: A Feasible Algorithm that Computes $[\underline{E}, \overline{E}]$ under Interval Uncertainty and Variance Constraint

- First, we compute the values

$$E^- \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} \underline{x}_i \text{ and } V^- \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (\underline{x}_i - E^-)^2;$$

$$E^+ \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i \text{ and } V^+ \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (\overline{x}_i - E^+)^2.$$

- If $V^- \le V_0$, then we return $\underline{E} = E^-$.
- If $V^+ \le V_0$, then we return $\overline{E} = E^+$.
- If $V_0 < V^-$ or $V_0 < V^+$, we sort the all $2n$ endpoints $\underline{x}_i$ and $\overline{x}_i$ into a non-decreasing sequence

$$z_1 \le z_2 \le \ldots \le z_{2n}$$

and consider $2n - 1$ *zones* $[z_k, z_{k+1}]$.

7. Algorithm (cont-d)

- For each zone $[z_k, z_{k+1}]$, we take:
  - for every i for which $\overline{x}_i \le z_k$, we take $x_i = \overline{x}_i$;
  - for every i for which $z_{k+1} \le \underline{x}_i$, we take $x_i = \underline{x}_i$;
  - for every other i, we take $x_i = \alpha$; let us denote the number of such i's by n_k.
- The value α is determined from the condition that for the selected vector x, we have $V(x) = V_0$:

$$\frac{1}{n} \cdot \left(\sum_{i:\overline{x}_i \le z_k} (\overline{x}_i)^2 + \sum_{i:z_{k+1} \le \underline{x}_i} (\underline{x}_i)^2 + n_k \cdot \alpha^2 \right) -$$

$$\frac{1}{n^2} \cdot \left(\sum_{i:\overline{x}_i \le z_k} \overline{x}_i + \sum_{i:z_{k+1} \le \underline{x}_i} \underline{x}_i + n_k \cdot \alpha \right)^2 = V_0.$$

8. Algorithm: Last Part

- If none of the two roots of the above quadratic equation belongs to the zone, this zone is dismissed.
- If one or more roots belong to the zone, then for each of these roots α, we compute the value

$$E_k(\alpha) = \frac{1}{n} \cdot \left(\sum_{i:\overline{x}_i \le z_k} \overline{x}_i + \sum_{i:z_{k+1} \le \underline{x}_i} \underline{x}_i + n_k \cdot \alpha \right).$$

- After that:
  - if $V_0 < V^-$, we return the smallest of the values $E_k(\alpha)$ as $\underline{E}$:
  $$\underline{E} = \min_{k,\alpha} E_k(\alpha);$$
  - if $V_0 < V^+$, we return the largest of the values $E_k(\alpha)$ as $\overline{E}$:
  $$\overline{E} = \max_{k,\alpha} E_k(\alpha).$$

9. Computation Time of the Algorithm

- Sorting $2n$ numbers requires time $O(n \cdot \log(n))$.
- Once the values are sorted, we can then go zone-by-zone, and perform the corresponding computations:
  - for each of $2n$ zones,
  - we compute several sums of n numbers.
- The sum for the first zone requires linear time.
- Once we have the sums for one zone, computing the sums for the next zone requires changing a few terms.
- Each value x_i changes status once, so overall, to compute all these sums, we need linear time $O(n)$.
- So, the total time is:

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

10. Toy Example

- Case: $n = 2$, $\mathbf{x}_1 = [-1, 0]$, $\mathbf{x}_2 = [0, 1]$, $V_0 = 0.16$.
- In this case, according to the above algorithm, we compute the values

$$E^- = \frac{1}{2} \cdot (-1 + 0) = -0.5; \quad E^+ = \frac{1}{2} \cdot (0 + 1) = 0.5;$$

$$V^- = \frac{1}{2} \cdot (((-1) - (-0.5))^2 + (0 - (-0.5))^2) = 0.25;$$

$$V^+ = \frac{1}{2} \cdot ((0 - 0.5)^2 + (1 - 0.5)^2) = 0.25.$$

- Here, $V_0 < V^-$ and $V_0 < V^+$, so we consider zones.
- By sorting the 4 endpoints -1, 0, 0, and 1, we get

$$z_1 = -1 \le z_2 = 0 \le z_3 = 0 \le z_4 = 1.$$

- Thus, here, we have three zones:

$$[z_1, z_2] = [-1, 0], \quad [z_2, z_3] = [0, 0], \quad [z_3, z_4] = [0, 1].$$

11. Toy Example (cont-d)

- *For the first zone* $[z_1, z_2] = [-1, 0]$, according to the above algorithm, we select $x_2 = 0$ and $x_1 = \alpha$, where
$$\frac{1}{2} \cdot (0^2 + \alpha^2) - \frac{1}{4} \cdot (0 + \alpha)^2 = V_0 = 0.16.$$
- Here, $\alpha = -0.8$ and $\alpha = 0.8$, and only the first root belongs to the zone $[-1, 0]$.
- For this root, we compute the value
$$E_1 = \frac{1}{2} \cdot (0 + \alpha) = \frac{1}{2} \cdot (0 + (-0.8)) = -0.4.$$
- *For the second zone* $[z_2, z_3] = [0, 0]$, according to the above algorithm, we select $x_1 = x_2 = 0$.
- In this case, there is no need to compute α, so we directly compute
$$E_2 = \frac{1}{2} \cdot (0 + 0) = 0.$$

12. Toy Example (end)

- *For the third zone* $[z_3, z_4] = [0, 1]$, according to the above algorithm, we select $x_1 = 0$ and $x_2 = \alpha$, where

$$\frac{1}{2} \cdot (0^2 + \alpha^2) - \frac{1}{4} \cdot (0 + \alpha)^2 = V_0 = 0.16.$$

- Of the two roots $\alpha = -0.8$ and $\alpha = 0.8$, only the second root belongs to the zone $[0, 1]$.

- For this root, we compute the value

$$E_3 = \frac{1}{2} \cdot (0 + \alpha) = \frac{1}{2} \cdot (0 + 0.8) = 0.4.$$

- *As a result*, we get the values E_k for all three zones; so, we return

$$\underline{E} = \min(E_1, E_2, E_3) = -0.4;$$

$$\overline{E} = \max(E_1, E_2, E_3) = 0.4.$$

13. Proof: Main Lemmas

- For $x'_i = -x_i$, we have $E' = -E$ and $V' = V$.
- Thus $\underline{E} = -\overline{E'}$; so, it is sufficient to consider $\overline{E}$.
- Let x be an optimizing vector, i.e., $E(x) = \overline{E}$.
- *Lemma 1:* if $x_i < E$, then $x_i = \overline{x}_i$.
- *Proof:* else, by adding $\Delta x_i > 0$ to x_i, we could increase E without increasing V.
- *Lemma 2:* if $\underline{x}_i < x_i < \overline{x}_i$, then:
  - for every j for which $E \le x_j < x_i$, we have $x_j = \overline{x}_j$;
  - for every k for which $x_k > x_i$, we have $x_k = \underline{x}_k$.
- *Proof:* similar.
- *Lemma 3:* if for all $x_i \ge E$, we have either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$, then $x_i = \overline{x}_i$ and $x_j = \underline{x}_j$ imply $x_i \le x_j$.

14. Proof (cont-d)

- *Lemma 1:* if $x_i < E$, then $x_i = \overline{x}_i$.
- *Lemma 2:* if $\underline{x}_i < x_i < \overline{x}_i$, then:
  - for every j for which $E \le x_j < x_i$, we have $x_j = \overline{x}_j$;
  - for every k for which $x_k > x_i$, we have $x_k = \underline{x}_k$.
- *Lemma 3:* if for all $x_i \ge E$, we have either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$, then $x_i = \overline{x}_i$ and $x_j = \underline{x}_j$ imply $x_i \le x_j$.
- Thus, there exists a threshold value α such that
  - for all j for which $x_j < \alpha$, we have $x_j = \overline{x}_j$;
  - for all k for which $x_k > \alpha$, we have $x_k = \underline{x}_k$.
- Once we know to which zone α belongs, we can uniquely determine all x_j of the corresponding vector x.
- Then $\overline{E}$ is the largest of the values $E(x)$ corresponding to different zones.