Estimating Variance under Interval and Fuzzy Uncertainty: Case of Hierarchical Estimation

Gang Xiang and Vladik Kreinovich

Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968, USA
email vladik@utep.edu

http://www.cs.utep.edu/vladik
http://www.cs.utep.edu/interval-comp

1. Estimating Variance under Uncertainty

- *Computing statistics is important:* traditional data processing starts with computing population mean and population variance:

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

- *Traditional approach:* assumes that we know the *exact* values x_i.
- *In practice:* these values come either from measurements or from expert estimates.
- *Uncertainty:* in both case, we get only *approximations* $\widetilde{x}_i$ to the actual (unknown) values x_i.
- *Result:* we only get approximate valued $\widetilde{E}$ and $\widetilde{V}$.
- *Question:* what is the accuracy of these approximations?

2. Case of Measurement Uncertainty

- The result $\widetilde{x}$ of the measurement is, in general, different from the (unknown) actual value x: $\Delta x \stackrel{\rm def}{=} \widetilde{x} - x \neq 0$.
- *Upper bound* Δ is usually supplied by the manufacturer: $|\Delta x| \le \Delta$.
- *Interval uncertainty:* $x \in [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.
- *Probabilistic approach:* often, we know probabilities of different values of Δx.
- *How these probabilities are determined:* by comparing with standard measuring instrument (SMI).
- *Cases when we do not know probabilities:*
  - cutting-edge measurements;
  - manufacturing.
- *Resulting problem:* find the ranges $\mathbf{E}$ and $\mathbf{V}$ of E and V.

3. Case of Expert Uncertainty

- *Situation:* an expert use natural language.
- *Example:* "most probably, the value of the quantity is between 6 and 7, but it is somewhat possible to have values between 5 and 8".
- *Natural formalization:* for every i, a fuzzy set $\mu_i(x_i)$.
- *Resulting problem:* given fuzzy numbers x_i, find the fuzzy numbers for E and V.
- *Reduction to interval case:* the α-cut for $C(x_1, \ldots, x_n)$ is equal to the range of C when x_i care in the corresponding α-cuts: $x_i \in \mathbf{x}_i(\alpha)$.
- *Conclusion:* for each characteristic $C(x_1, \ldots, x_n)$, it is sufficient to be able to compute the range

$$C(\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{\text{def}}{=} \{C(x_1, \ldots, x_n) \mid x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

4. Estimating Mean under Interval Uncertainty: What Is Known

- *Fact:* the arithmetic average $E(x_1, \ldots, x_n)$ is an increasing function of $x_1, \ldots, x_n$.
- *Conclusions:*
  - the smallest possible value $\underline{E}$ of E is attained when each value x_i is the smallest possible ($x_i = \underline{x}_i$);
  - the largest possible value $\overline{E}$ of E is attained when $x_i = \overline{x}_i$ for all i.
- *Resulting formulas:* the range $\mathbf{E}$ of E is equal to

$$[E(\underline{x}_1, \ldots, \underline{x}_n), E(\overline{x}_1, \ldots, \overline{x}_n)],$$

i.e., to

$$\mathbf{E} = [\underline{E}, \overline{E}] = \left[\frac{1}{n} \cdot (\underline{x}_1 + \ldots + \underline{x}_n), \frac{1}{n} \cdot (\overline{x}_1 + \ldots + \overline{x}_n)\right].$$

5. Estimating Variance under Interval Uncertainty: What is Known

- *Problem:* compute the range $\mathbf{V} = [\underline{V}, \overline{V}]$ of the variance V over interval data $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.
- *Known:* there is a polynomial-time algorithm for computing $\underline{V}$.
- *In general:* computing $\overline{V}$ is NP-hard.
- *In many practical situations:* there are efficient algorithms for computing $\overline{V}$.
- *Example:* consider narrowed intervals $[x_i^-, x_i^+]$, where $x_i^- \stackrel{\text{def}}{=} \widetilde{x}_i - \dfrac{\Delta_i}{n}$ and $x_i^+ \stackrel{\text{def}}{=} \widetilde{x}_i + \dfrac{\Delta_i}{n}$.
- *Case:* no two narrowed intervals are proper subsets of one another, i.e., $[x_i^-, x_i^+] \not\subseteq (x_j^-, x_j^+)$ for all i and j.
- *For this case:* here exists an $O(n \cdot \log(n))$ time algorithm for computing $\overline{V}$.

6. Hierarchical Case: Formulation of the Problem

- *Situation:* often,
  - we do not know the *individual* values of the observations x_i,
  - we only have *average* values corresponding to several ($m < n$) groups $I_1, \ldots, I_m$ of observations.
- *Typically:* for each group j, we know
  - the *frequency* p_j of this group (i.e., the probability that a randomly selected observation belongs to this group),
  - the *average* E_j over this group, and
  - the *standard deviation* σ_j within j-th group.
- *Formulas:* $E = \sum\limits_{j=1}^{m} p_j \cdot E_j$ and $V = V_E + V_\sigma$, where

$$V_E \stackrel{\text{def}}{=} M_E - E^2,\ M_E \stackrel{\text{def}}{=} \sum_{j=1}^{m} p_j \cdot E_j^2, \text{ and } V_\sigma \stackrel{\text{def}}{=} \sum_{j=1}^{m} p_j \cdot \sigma_j^2.$$

7. Hierarchical Case: Interval Uncertainty

- *Practical situation:* we only know the intervals $\mathbf{E}_j = [\underline{E}_j, \overline{E}_j]$ and $[\underline{\sigma}_j, \overline{\sigma}_j]$ that contain E_j and σ_j.
- *Mean* E is monotonic in E_j, hence

$$\mathbf{E} = [\underline{E}, \overline{E}] = \left[\sum_{j=1}^{m} p_j \cdot \underline{E}_j, \sum_{j=1}^{m} p_j \cdot \overline{E}_j\right].$$

- *Variance:* the terms V_E and V_σ in the expression for V depend on different variables.
- *Conclusion:* the range $\mathbf{V} = [\underline{V}, \overline{V}]$ of the population variance V is equal to the sum of the ranges $\mathbf{V}_E = [\underline{V}_E, \overline{V}_E]$ and $\mathbf{V}_\sigma = [\underline{V}_\sigma, \overline{V}_\sigma]$.
- Due to monotonicity, $\mathbf{V}_\sigma = \left[\sum\limits_{j=1}^{m} p_j \cdot (\underline{\sigma}_j)^2, \sum\limits_{j=1}^{m} p_j \cdot (\overline{\sigma}_j)^2\right]$.
- Thus, it is sufficient to compute $\mathbf{V}_E$.

8. Formulation of the Problem in Precise Terms

GIVEN:

- an integer $m \geq 1$;
- m numbers $p_j > 0$ for which $\sum\limits_{j=1}^{m} p_j = 1$; and
- m intervals $\mathbf{E}_j = [\underline{E}_j, \overline{E}_j]$.

COMPUTE the range

$$\mathbf{V}_E = \{V_E(E_1, \ldots, E_m) \mid E_1 \in \mathbf{E}_1, \ldots, E_m \in \mathbf{E}_m\},$$

where

$$V_E \stackrel{\text{def}}{=} \sum_{j=1}^{m} p_j \cdot E_j^2 - E^2; \quad E \stackrel{\text{def}}{=} \sum_{j=1}^{m} p_j \cdot E_j.$$

9. Analysis of the Problem

- *Fact:* the function V_E is convex.
- *Fact:* the box $\mathbf{E}_1 \times \ldots \times \mathbf{E}_m$ is convex.
- *Known:* a polynomial-time algorithm for computing minima of convex functions on convex sets.
- *Conclusion:* we can compute $\underline{V}_E$ in polynomial time.
- *Computing* $\overline{V}_E$*:* in general, NP-hard.
- *Proof of NP-hardness:*
  - for $p_1 = \ldots = p_m = \dfrac{1}{m}$, the expression V_E becomes a standard formula for the sample variance V;
  - so, in this case, $\overline{V}_E = \overline{V}$;
  - computing $\overline{V}$ under interval uncertainty is NP-hard;
  - thus, the more general problem of computing $\overline{V}_E$ is also NP-hard.

10. Efficient Algorithm for Computing $\overline{V}_E$

- *Notations:* $\widetilde{E}_j \stackrel{\text{def}}{=} \dfrac{\underline{E}_j + \overline{E}_j}{2}$, $\Delta_j \stackrel{\text{def}}{=} \dfrac{\overline{E}_j - \underline{E}_j}{2}$.
- *Narrowed intervals* $[E_j^-, E_j^+]$, where $E_j^- \stackrel{\text{def}}{=} \widetilde{E}_j - p_j \cdot \Delta_j$ and $E_j^+ \stackrel{\text{def}}{=} \widetilde{E}_j + p_j \cdot \Delta_j$.
- *Case:* no two narrowed intervals are proper subsets of each other, i.e., $[E_j^-, E_j^+] \not\subseteq (E_k^-, E_k^+)$ for all j and k.
- *Efficient $O(m \cdot \log(m))$ algorithm for this case:*
  - First, sort the $\widetilde{E}_1, \ldots, \widetilde{E}_m$ into an increasing sequence $\widetilde{E}_1 \le \widetilde{E}_2 \le \ldots \le \widetilde{E}_m$.
  - Then, for every k from 0 to m, compute the value $V_E^{(k)} = M^{(k)} - (E^{(k)})^2$ of V_E for the vector $\vec{E}^{(k)} = (\underline{E}_1, \ldots, \underline{E}_k, \overline{E}_{k+1}, \ldots, \overline{E}_m)$.
  - Finally, compute $\overline{V}_E$ as the largest of $m+1$ values $V_E^{(0)}, \ldots, V_E^{(m)}$.

11. Number of Computation Steps

- *Known:* sorting requires $O(m \cdot \log(m))$ steps.
- Computing the initial values $M^{(0)}$, $E^{(0)}$, and $V_E^{(0)}$ requires linear time $O(m)$.
- *Reminder:* $V_E^{(k)} = M^{(k)} - (E^{(k)})^2$ is the value for the vector $\vec{E}^{(k)} = (\underline{E}_1, \ldots, \underline{E}_k, \overline{E}_{k+1}, \ldots, \overline{E}_m)$.
- *Transition:* once we have $M^{(k)} = \sum\limits_{j=1}^{m} p_j \cdot \left(E_j^{(k)}\right)^2$ and $E^{(k)} = \sum\limits_{j=1}^{m} p_j \cdot E_j^{(k)}$, we compute, in $O(1)$ steps,
$$M^{(k+1)} = M^{(k)} + p_{k+1} \cdot (\underline{E}_{k+1})^2 - p_{k+1} \cdot (\overline{E}_{k+1})^2,$$
$$E^{(k+1)} = E^{(k)} + p_{k+1} \cdot \underline{E}_{k+1} - p_{k+1} \cdot \overline{E}_{k+1}.$$
- Finding the largest of $V_E^{(0)}, \ldots, V_E^{(m)}$ requires $O(m)$ steps.
- Thus, overall, we need $O(m \cdot \log(m)) + O(m) + m \cdot O(1) + O(m) = O(m \cdot \log(m))$ steps.

12. Acknowledgments

This work was supported in part:

- by NSF grants EAR-0225670 and DMS-0532645 and
- by the Texas Department of Transportation grant No. 0-5453.