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

Ali Jalal-Kamali

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

1. Need for Correlation

- In practice, it is often desirable to know which quantities x, y are independent and which are correlated.
- To estimate the correlation ρ between x and y, we measure the values x_i and y_i in different situations i.
- ρ is then estimated as the ratio $\rho = \dfrac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}$, where the covariance C and variances V_x, V_y are:

$$C \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - E_x \cdot E_y,$$

$$V_x \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x)^2, \quad V_y \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - E_y)^2, \ \text{and}$$

$$E_x \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \quad E_y \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} y_i.$$

2. Need to Take into Account Interval Uncertainty

- The values x_i and y_i used to estimate correlation come from measurements.
- Measurements are never absolutely accurate.
- The measurement results $\widetilde{x}_i$ and $\widetilde{y}_i$ are, in general, different from the actual (unknown) values x_i and y_i.
- Hence, the value $\widetilde{\rho}$ based on $\widetilde{x}_i$ and $\widetilde{y}_i$ is, in general, different from the ideal value ρ based on x_i and y_i.
- It is therefore desirable to determine how accurate is the resulting estimate.
- Sometimes, we know the probabilities of different values of $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ and $\Delta y_i \stackrel{\text{def}}{=} \widetilde{y}_i - y_i$.
- However, in many cases, we do not know these probabilities.

3. Interval Uncertainty (cont-d)

- In many cases, we do not know the probabilities of different values Δx_i and Δy_i.
- We only know the upper bounds Δ_{xi} and Δ_{yi} on the corresponding measurement errors:

$$|\Delta x_i| \le \Delta_{xi} \text{ and } |\Delta y_i| \le \Delta_{yi}.$$

- In this case, the only info that we have about x_i and y_i is that they belong to the intervals

$$[\underline{x}_i, \overline{x}_i] = [\widetilde{x}_i - \Delta_{xi}, \widetilde{x}_i + \Delta_{xi}] \text{ and } [\underline{y}_i, \overline{y}_i] = [\widetilde{y}_i - \Delta_{yi}, \widetilde{y}_i + \Delta_{yi}].$$

- Different values $x_i \in [\underline{x}_i, \overline{x}_i]$ and $y_i \in [\underline{y}_i, \overline{y}_i]$ lead, in general, to different values of the correlation.
- It is therefore desirable to find the range $[\underline{\rho}, \overline{\rho}]$ of all possible values of the correlation ρ:

$$\{\rho(x_1, \ldots, x_n, y_1, \ldots, y_n) : x_i \in [\underline{x}_i, \overline{x}_i], y_i \in [\underline{y}_i, \overline{y}_i]\}.$$

4. Expert Uncertainty Reduced to the Interval Uncertainty

- An expert usually describes his/her uncertainty by using words from the natural language.
- To formalize this knowledge, fuzzy set theory is used, in which
  - for every quantity x_i, we have a fuzzy set $\mu_i(x_i)$,
  - which describes the expert's knowledge about x_i.
- An alternative user-friendly way to represent a fuzzy set is by using its α-cuts $\mathbf{x}_i(\alpha) \stackrel{\text{def}}{=} \{x_i : \mu(x_i) \geq \alpha\}$.
- It is known that for any function $y = f(x_1, \ldots, x_n)$, the α-cut of y is equal to

$$\mathbf{y}(\alpha) = \{f(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1(\alpha), \ldots, x_n \in \mathbf{x}_n(\alpha)\}.$$

- So, estimating ρ under fuzzy uncertainty can be reduced to interval uncertainty.

5. What Is Known

- Estimating correlation under interval uncertainty is, in general, NP-hard.
- Unless P=NP, there is no feasible algorithm for computing the range of correlation.
- It is known that:
  - while we cannot have an efficient algorithm for computing both bounds $\underline{\rho}$ and $\overline{\rho}$,
  - we can effectively compute (at least) one of the bounds.
- We can effectively compute $\overline{\rho}$ when $\overline{\rho} > 0$ and we can effectively compute $\underline{\rho}$ when $\underline{\rho} < 0$.
- Eff. comp. are also possible for *weighted* correlation, w/$E_x = \sum\limits_{i=1}^{n} w_i \cdot x_i$, etc., for some $w_i \geq 0$ s.t. $\sum\limits_{i=1}^{n} w_i = 1$.

6. Estimation Is Usually Hierarchical

- In some practical situations, e.g., when processing census results, we do not process all of the data at once:
  - we first combine the data by county,
  - then combine county data into state-wide data, etc.
- In general, in each stage, the data points are divided into groups $I_1, \ldots, I_m$; e.g., the overall average E_x is:

$$E_x = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot \sum_{j=1}^{m} \sum_{i \in I_j} x_i = \sum_{j=1}^{m} p_j \cdot E_{xj},$$

$$\text{where } E_{xj} = \frac{1}{n_j} \cdot \sum_{i \in I_j} x_i \text{ and } p_j \stackrel{\text{def}}{=} \frac{n_j}{n}.$$

- We compute E_{xj} for each group and then compute E_x.
- Similarly, $E_y = \sum\limits_{j=1}^{m} p_j \cdot E_{yj}$.

7. Estimation Is Usually Hierarchical (cont-d)

- *Reminder:* $E_x = \sum_{j=1}^{m} p_j \cdot E_{xj}$ and $E_y = \sum_{j=1}^{m} p_j \cdot E_{yj}$.
- Similarly, $V_x = \sum_{j=1}^{m} p_j \cdot (E_{xj} - E_x)^2 + \sum_{j=1}^{m} p_j \cdot V_{xj}$, where V_{xj} are x-variances within the j-th group.
- Also, $V_y = \sum_{j=1}^{m} p_j \cdot (E_{yj} - E_y)^2 + \sum_{j=1}^{m} p_j \cdot V_{yj}$, where V_{xj} are y-variances within the j-th group.
- Cov. $C = \sum_{j=1}^{m} p_j \cdot (E_{xj} - E_x) \cdot (E_{yj} - E_y) + \sum_{j=1}^{m} p_j \cdot C_j$, where C_j is the covariance over the j-th group.
- Finally, we compute correlation ρ as

$$\rho = \frac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}.$$

8. Hierarchical Estimation Under Interval Uncertainty

- Ideally, for each group j, we compute the values p_j, E_{xj}, E_{yj}, V_{xj}, V_{yj}, and C_j.
- Based on these values, we compute E, V_x, V_y, C, ρ.
- In practice, we often only know the values x_i and y_i with interval uncertainty.
- As a result, for each group j, we only know the interval of possible values for each characteristic.
- That means that we only know the intervals $\mathbf{E}_{xj}$, $\mathbf{E}_{xj}$, $\mathbf{E}_{yj}$, $\mathbf{V}_{xj}$, $\mathbf{V}_{yj}$, and $\mathbf{C}_j$.
- Different values from these intervals lead to different ρ.
- It is desirable to find the range $[\underline{\rho}, \overline{\rho}]$.
- We show that for hierarchical estimation, it is feasible to compute at least one of the endpoints of $[\underline{\rho}, \overline{\rho}]$.

9. Main Result

- There exists a polynomial-time algorithm that:
  - given intervals $\mathbf{E}_{xj}$, $\mathbf{E}_{xj}$, $\mathbf{E}_{yj}$, $\mathbf{V}_{xj}$, $\mathbf{V}_{yj}$, and $\mathbf{C}_j$,
  - computes (at least) one of the endpoint of the interval $[\underline{\rho}, \overline{\rho}]$ of possible values of the correlation ρ.
- Specifically, in the case of a non-degenerate interval $[\underline{\rho}, \overline{\rho}]$:
  - when $\overline{\rho} \le 0$, we compute the lower endpoint $\underline{\rho}$;
  - when $0 \le \underline{\rho}$, we compute the upper endpoint $\overline{\rho}$;
  - in all remaining cases, we compute both endpoints $\underline{\rho}$ and $\overline{\rho}$.

10. Reducing Minimum to Maximum

- When we change the sign of y_i, the correlation changes sign as well:

$$\rho(x_1, \ldots, x_n, -y_1, \ldots, -y_n) = -\rho(x_1, \ldots, x_n, y_1, \ldots, y_n).$$

- If z goes from $\underline{z}$ to $\overline{z}$, the range of $-z$ is $[-\overline{z}, -\underline{z}]$.
- So, for the endpoints of the ranges, we get

$$\overline{\rho}([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n], -[\underline{y}_1, \overline{y}_1], \ldots, -[\underline{y}_n, \overline{y}_n]) =$$

$$-\underline{\rho}([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n], [\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]),$$

$$\text{where } -[\underline{y}_i, \overline{y}_i] = \{-y_i : y_i \in [\underline{y}_i, \overline{y}_i]\} = [-\overline{y}_i, -\underline{y}_i].$$

- If we know how to compute $\overline{\rho}$, we can compute $\underline{\rho}$ as

$$\underline{\rho}([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n], [\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]) =$$

$$-\overline{\rho}([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n], [-\overline{y}_1, -\underline{y}_1], \ldots, [-\overline{y}_n, -\underline{y}_n]).$$

- Thus, we can concentrate on computing $\overline{\rho}$.

11. Preliminary Observation

- *Reminder:* $\rho = \dfrac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}$.
- In the ratio ρ:
  - the dependence on C_j is only in the numerator C;
  - the dependence on V_{xj} and V_{yj} is only in the denominator $\sqrt{V_x} \cdot \sqrt{V_y}$.
- Thus, the ratio ρ is the largest when:
  - each term C_j attains its largest possible value $\overline{C}_j$;
  - each term V_{xj} and V_{yj} attains its smallest possible value $\underline{V}_{xj}$ and $\underline{V}_{yj}$.
- So, in the following text:
  - we will take $C_j = \overline{C}_j$, $V_{xj} = \underline{V}_{xj}$, and $V_{yj} = \underline{V}_{yj}$, and
  - consider only the dependence on E_{xj} and E_{yj}.

12. Algorithm

- For each j from 1 to m, the box $[\underline{E}_{xj}, \overline{E}_{xj}] \times [\underline{E}_{yj}, \overline{E}_{yj}]$ has four vertices:

$$(\underline{E}_{xj}, \underline{E}_{yj}), \quad (\underline{E}_{xj}, \overline{E}_{yj}), \quad (\overline{E}_{xj}, \underline{E}_{yj}), \quad (\overline{E}_{xj}, \overline{E}_{yj}).$$

- Let's consider 4-tuples consisting of two vertices and two signs $(-,-)$, $(-,0)$, $\ldots$, $(+,+)$.
- For the first vertex, we:
  - slightly increase x if the first sign is $+$ and
  - slightly decrease x if the first sign is $-$.
- We similarly move the second vertex depending on the second sign.
- We form a straight line through the resulting points.
- We select two 4-tuples, and form two lines: *representative x-line* and *representative y-line.*

13. Algorithm (cont-d)

- We have an actual x-line $y = E_y + k_x \cdot (x - E_x)$ and an actual y-line $x = E_x + k_y \cdot (y - E_y)$.
- Here, E_x, E_y, k_x, k_y are to-be-determined.
- For each box, based on its location in comparison to the representative lines, we select E_{xj} and E_{yj}:
- If the box is above the repr. x-line, take $E_{xj} = \overline{E}_{xj}$.
- Pick E_{yj} s.t. $(\overline{E}_{xj}, E_{yj})$ is closest to the actual y-line.
- If the box is below the x-line, we take $E_{xj} = \underline{E}_{xj}$.
- If the box is to the right of the y-line, take $E_{yj} = \underline{E}_{yj}$.
- Pick E_{xj} s.t. $(E_{xj}, \underline{E}_{yj})$ is closest to the actual x-line.
- If the box is left of the repr. y-line, take $E_{yj} = \overline{E}_{yj}$.
- When the box contains the intersection point (E_x, E_y) of x- and y-lines, take $E_{xj} = E_x$ and $E_{yj} = E_y$.

14. Algorithm (cont-d)

- For each i, we get explicit expressions for E_{xj} and E_{yj} in terms of the four unknowns E_x, E_y, k_x and k_y.
- By substituting these expressions into the following formulas, we get a system of 4 equations with 4 unknowns:

$$E_x = \sum_{j=1}^{m} p_j \cdot E_{xj}; \quad E_y = \sum_{j=1}^{m} p_j \cdot E_{yj};$$

$$\sum_{j=1}^{m} p_j \cdot E_{xj} \cdot E_{yj} - E_x \cdot E_y + \sum_{j=1}^{m} p_j \cdot \overline{C}_j =$$

$$k_x \cdot \left(\sum_{j=1}^{m} p_j \cdot (E_{xj} - E_x)^2 + \sum_{j=1}^{m} p_j \cdot \underline{V}_{xj} \right) =$$

$$k_y \cdot \left(\sum_{j=1}^{m} p_j \cdot (E_{yj} - E_y)^2 + \sum_{j=1}^{m} p_j \cdot \underline{V}_{yj} \right).$$

15. Algorithm (final part)

- We solve the system of 4 equations with 4 unknowns:

$$E_x = \sum_{j=1}^{m} p_j \cdot E_{xj}; \quad E_y = \sum_{j=1}^{m} p_j \cdot E_{yj};$$

$$\sum_{j=1}^{m} p_j \cdot E_{xj} \cdot E_{yj} - E_x \cdot E_y + \sum_{j=1}^{m} p_j \cdot \overline{C}_j =$$

$$k_x \cdot \left(\sum_{j=1}^{m} p_j \cdot (E_{xj} - E_x)^2 + \sum_{j=1}^{m} p_j \cdot \underline{V}_{xj} \right) =$$

$$k_y \cdot \left(\sum_{j=1}^{m} p_j \cdot (E_{yj} - E_y)^2 + \sum_{j=1}^{m} p_j \cdot \underline{V}_{yj} \right).$$

- For each of the solutions E_x, E_y, k_x and k_y, we compute E_{xj} and E_{yj} ($j = 1, \ldots, m$), and then the correlation ρ.
- The largest of these values ρ is returned as $\overline{\rho}$.

16. Computation Time

- We have $4m$ possible vertices, so we have $O(m^2)$ possible pairs of vertices – hence $O(m^2)$ possible 4-tuples.
- Thus, we have $O(m^2)$ possible representative x-lines, and we also have $O(m^2)$ representative y-lines.
- In our algorithms, we consider pairs consisting of a representative x-line and a representative y-line.
- We have $O(m^2)\cdot O(m^2) = O(m^4)$ possible pairs of lines.
- For each pair of lines, we need:
  - $O(m)$ steps to select E_{xj}, E_{yj} for each of m boxes;
  - $O(m)$ steps to compute ρ;
  - to the total of $O(m) + O(m) = O(m)$.
- Thus, the total computation time is $O(m^4) \times O(m) = O(m^5)$, which is polynomial (feasible).

17. Towards Proving the Result: Reminder

- A function $f(x)$ defined on an interval $[\underline{x}, \overline{x}]$ attains its minimum:
  - either an internal point $x \in (\underline{x}, \overline{x})$,
  - or at one of its endpoints $x = \underline{x}$ or $x = \overline{x}$.
- If the minimum of $f(x)$ is attained at an internal point, then
$$\frac{df}{dx} = 0.$$
- If the minimum is attained for $x = \underline{x}$, then
$$\frac{df}{dx} \geq 0.$$
- If the minimum is attained for $x = \overline{x}$, then
$$\frac{df}{dx} \leq 0.$$

18. Proof of the Result

- $$\frac{\partial \rho}{\partial E_{xj}} = \frac{1}{\sigma_x \cdot \sigma_y \cdot n} \cdot [(E_{yj} - E_y) - k_x \cdot (E_{xj} - E_x)],\ k_x = \frac{C}{V_x}.$$
- Thus, the sign of the derivative coincides with the sign of the expression $(E_{yj} - E_y) - k_x \cdot (E_{xj} - E_x)$.
- So, the sign depends on whether we are above or below the actual x-line $E_{yj} = E_y + k_x \cdot (E_{xj} - E_x)$.
- The sign of $\dfrac{\partial \rho}{\partial E_{yj}}$ depends on where we are w.r.t. the actual y-line $E_{xj} = E_x + k_y \cdot (E_{yj} - E_y)$, with $k_y = \dfrac{C}{V_y}$.
- Now, the selection of E_{xj} and E_{yj} follows from calculus.
- All possible locations of lines w.r.t. vertices are covered:
  - each line can be moved and rotated
  - until it almost touches two points – i.e., becomes one of our representative lines.

19. Acknowledgments

The author is thankful to Prof. Vladik Kreinovich for all the help and guidance.