Estimating Covariance for Privacy Case under Interval and Fuzzy Uncertainty

Ali Jalal-Kamali, Vladik Kreinovich, and Luc Longpré

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

1. Need to preserve privacy in statistical databases

- In order to find relations between different quantities, we *collect* a large amount of *data*.
- *Example:* we collect *medical* data to try to find correlations between a disease and lifestyle factors.
- In some cases, we are looking for commonsense correlations, e.g., between smoking and lung diseases.
- For statistical databases to be most useful, we need to *allow researchers* to *ask* arbitrary *questions*.
- However, this may inadvertently *disclose* some *private information* about the individuals.
- Therefore, it is desirable to *preserve privacy* in statistical databases.

2. Intervals as a way to preserve privacy in statistical databases

- One way to preserve privacy is to store *ranges* (intervals) rather than the exact data values.
- This makes sense from the viewpoint of a statistical database.
- In general, this is how data is often collected:
  - we set some *threshold* values $t_0, \ldots, t_N$ and
  - ask a person whether the actual value x_i is in the interval $[t_0, t_1]$, or $\ldots$, or in the interval $[t_{N-1}, t_N]$.
- As a result, for each quantity x and for each person i:
  - instead of the *exact* value x_i,
  - we store an *interval* $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ that contains x_i.
- Each of these intervals coincides with one of the given ranges $[t_0, t_1]$, $[t_1, t_2]$, $\ldots$, $[t_{N-1}, t_N]$.

3. Need to estimate covariance and correlation under interval uncertainty

- One of the main objectives of collecting data is to find *correlations* between different variables.
- A correlation $\rho_{x,y}$ between two quantities x and y is defined as: $\rho_{x,y} = \dfrac{C_{x,y}}{\sigma_x \cdot \sigma_y}$; $\sigma_x = \sqrt{V_x}$, $\sigma_y = \sqrt{V_y}$,

$$C_{x,y} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - E_x \cdot E_y$$

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

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

- So, we need to find the *range* of $C_{x,y}(x_1, \ldots, x_n, y_1, \ldots, y_n)$.

4. Estimating statistical characteristics under interval uncertainty: what is known

- General problem of *interval computations*: estimating the range

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

  – of a given function $f(x_1, \ldots, x_n)$

  – on given intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- The need for interval computations comes beyond privacy concerns.
- Usually, data come from measurements, and measurements are never absolutely accurate.
- Often, the only information about the measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ is the upper bound Δ_i: $|\Delta x_i| \le \Delta_i$.
- So, the actual value x_i is in the interval

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

5. Estimating statistical characteristics for privacy case under interval uncertainty

- *What is known:*
  - for the general case,
  - the problems of computing the range of variance and covariance are NP-hard.
- *What is known:*
  - for privacy case,
  - the range of *variance* can be computed in polynomial time.
- *In this paper we show that:*
  - for privacy case,
  - the range of *covariance* can also be computed in polynomial time.

6. Possibility of extending our results of the fuzzy case

- An alternative way to preserve privacy is to have fuzzy thresholds.
- This possibility goes beyond privacy preservation.
- We can provide reasonable estimates in terms of words from natural language. In this case,
  - for each i, instead of an interval $\mathbf{x}_i$,
  - we have a fuzzy number X_i describing the corr. natural language word, with a membership f-n $\mu_i(x_i)$.
- For $C(x_1, \ldots, x_n)$, Zadeh's extension principle defines, for fuzzy inputs $X_1, \ldots, X_n$, the fuzzy value

$$Y = C(X_1, \ldots, X_n).$$

7. Possibility of extending our results of the fuzzy case (cont-d)

- Zadeh's extension can be expressed in terms of α-cuts

  $$X_i(\alpha) \stackrel{\text{def}}{=} \{x_i : \mu_i(x_i) \ge \alpha\} \text{ and } C(\alpha) \stackrel{\text{def}}{=} \{y : \mu(y) \ge \alpha\}.$$

- Specifically, for every α:

  $$C(\alpha) = \{C(x_1, \ldots, x_n) : x_1 \in X_1(\alpha), \ldots, x_n \in X_n(\alpha)\}.$$

- Thus, for each $\alpha \in (0, 1]$:
  - the corresponding α-cut $C(\alpha)$
  - can be obtained by solving the corresponding interval computations problem.
- Therefore, in the following paper, we only consider the case of interval uncertainty.

8. Formulation of the problem

- *Given:*
  - x-thresholds $t_0^{(x)}, t_1^{(x)}, \ldots, t_{N_x}^{(x)}$;
  - y-thresholds $t_0^{(y)}, t_1^{(y)}, \ldots, t_{N_y}^{(y)}$;
  - n pairs of intervals $(\mathbf{x}_i, \mathbf{y}_i)$ in which:
    - each of $\mathbf{x}_i$ is one of the x-ranges $[t_k^{(x)}, t_{k+1}^{(x)}]$, and
    - each of $\mathbf{y}_i$ is one of the y-ranges $[t_\ell^{(y)}, t_{\ell+1}^{(y)}]$.
- *Compute:* the range $[\underline{C}_{x,y}, \overline{C}_{x,y}]$ of possible values of

$$C_{x,y} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - E_x \cdot E_y,$$

where

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

9. Reducing computing $\overline{C}_{x,y}$ to computing $\underline{C}_{x,y}$

- We need to compute both the maximum $\overline{C}_{x,y}$ and the minimum $\underline{C}_{x,y}$.
- When we change the sign of y_i, the covariance changes sign as well: $C_{xy}(x_i, -y_i) = -C_{xy}(x_i, y_i)$.
- Thus, for the ranges, we get $\mathbf{C}_{xy}(\mathbf{x}_i, -\mathbf{y}_i) = -\mathbf{C}_{xy}(\mathbf{x}_i, \mathbf{y}_i)$.
- Since the function $z \to -z$ is decreasing:
  - its smallest value is attained when z is the largest;
  - its largest value is attained when z is the smallest.
- Thus, if z goes from $\underline{z}$ to $\overline{z}$, the range of $-z$ is $[-\overline{z}, -\underline{z}]$.
- Therefore, $\underline{C}_{xy}(\mathbf{x}_i, -\mathbf{y}_i) = -\overline{C}_{xy}(\mathbf{x}_i, \mathbf{y}_i)$.
- Thus, if we know how to compute $\underline{C}_{xy}(\mathbf{x}_i, \mathbf{y}_i)$, we can then compute $\overline{C}_{xy}(\mathbf{x}_i, \mathbf{y}_i)$ as $\overline{C}_{xy}(\mathbf{x}_i, \mathbf{y}_i) = -\underline{C}_{xy}(\mathbf{x}_i, -\mathbf{y}_i)$.
- So, we will now only talk about computing $\underline{C}_{x,y}$.

10. Algorithm for computing $\underline{C}_{xy}$: main idea

- We have N_x possible x-ranges $[t_k^{(x)}, t_{k+1}^{(x)}]$.
- We also have N_y possible y-ranges $[t_\ell^{(y)}, t_{\ell+1}^{(x)}]$.
- So, totally, we have $N_x \cdot N_y$ cells $[t_k^{(x)}, t_{k+1}^{(x)}] \times [t_\ell^{(y)}, t_{\ell+1}^{(y)}]$.
- In this algorithm, we analyze these cells c one by one.
- For each c, we assume that the pair (E_x, E_y) corresponding to the minimizing set (x_i, y_i) is contained in c.
- We then find the values (x_i, y_i) where, under this assumption, the minimum of C_{xy} is attained.
- Based on these values x_i and y_i, we compute E_x, E_y.
- If $(E_x, E_y) \in c$, we compute the value C_{xy}.
- The smallest of the corresponding values C_{xy} is the desired minimum $\underline{C}_{xy}$.

11. Possible position of intervals $\mathbf{x}_i$ and $\mathbf{y}_i$ in relation to the cell

- For each cell $[t_k^{(x)}, t_{k+1}^{(x)}] \times [t_\ell^{(y)}, t_{\ell+1}^{(y)}]$ and for each i, there are three possible positions for $\mathbf{x}_i$:

  X^0: $\mathbf{x}_i$ coincides with the cell's x-range;
  X^-: $\mathbf{x}_i$ is to the left of the x-range;
  X^+: $\mathbf{x}_i$ is to the right of the x-range.

- Similarly, there are three possible positions for $\mathbf{y}_i$:

  Y^0: $\mathbf{y}_i$ coincides with the cell's y-range;
  Y^-: $\mathbf{y}_i$ is to the left of the y-range;
  Y^+: $\mathbf{y}_i$ is to the right of the y-range.

- So, we have $3 \cdot 3 = 9$ pairs of options.

12. Selecting x_i and y_i at which C_{xy} attains its minimum

For each cell c and for each i, the minimum of $\underline{C}_{xy}$ under the assumption $(E_x, E_y) \in c$ is attained:

- in case (X^+, Y^+): for $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$;
- in case (X^+, Y^0): for $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$;
- in case (X^+, Y^-): for $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$;
- in case (X^-, Y^+): for $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$;
- in case (X^-, Y^0): for $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$;
- in case (X^-, Y^-): for $x_i = \overline{x}_i$ and $y_i = \overline{y}_i$;
- in case (X^0, Y^+): for $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$;
- in case (X^0, Y^-): for $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$;
- in case (X^0, Y^0): for $(x_i, y_i) = (\underline{x}_i, \underline{y}_i)$ or for $(x_i, y_i) = (\overline{x}_i, \overline{y}_i)$.

13. Implementation details

- For those i for which $\mathbf{x}_i \times \mathbf{y}_i \neq c$, we directly compute the minimizing values x_i and y_i.
- For each i for which $\mathbf{x}_i \times \mathbf{y}_i = c$, we have two different options: $(x_i, y_i) = (\underline{x}_i, \underline{y}_i)$ and $(x_i, y_i) = (\overline{x}_i, \overline{y}_i)$.
- A naive implementation would require testing all 2^M combinations, where M is the number of such cells.
- Luckily, the value C_{xy} does not change if we swap pairs (x_i, y_i).
- So, the value C_{xy} only depends on the number of i's to which we assign $(x_i, y_i) = (\underline{x}_i, \underline{y}_i)$.
- Thus, we can make computations efficient if, for each integer $m = 0, 1, 2, \ldots, M$, we assign:
  - to m i's, the values $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$, and
  - to the rest, the values $x_i = \overline{x}_i$ and $y_i = \overline{y}_i$.

14. Resulting computation time of our algorithm

- For each cell, we perform $M+1 \leq n$ computations C_{xy} – one for each option m.
- In general, computing $E_x = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$, $E_y = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$, and $C_{x,y} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y)$ takes time $O(n)$.
- However, each new computation differs from the previous one
  - by a single change in $\sum x_i \cdot y_i$ and
  - a single change in estimating $E_x \sim \sum x_i$ and $E_y \sim \sum y_i$.
- Thus, each new computation requires $O(1)$, and so, for each cell, the total computation time is $O(n)$.
- So, for all $N_x \cdot N_y$ cells, we need time $O(N_x \cdot N_y \cdot n)$.

15. Computation time: discussion

- *Reminder:* this algorithm takes time $O(N_x \cdot N_y \cdot n)$.
- Usually, the number N_x of x-ranges and the number N_y of y-ranges are fixed.
- In this case, what we have is a *linear-time* algorithm.
- Clearly, it is not possible to compute covariance faster than in linear time:
  - we need to take into account all n data points, and
  - processing each data point requires at least one computation.
- So, our algorithm is *(asymptotically) optimal* – it requires the smallest possible order of computation time $O(n)$.
- *Comment:* for general (non-privacy) intervals, the problem is NP-hard.

16. Computing $\overline{C}_{xy}$

- We use the fact that $\overline{C}_{xy}(\mathbf{x}_i, \mathbf{y}_i) = -\underline{C}_{xy}(\mathbf{x}_i, -\mathbf{y}_i)$.
- We form N_y threshold values for $z \stackrel{\text{def}}{=} -y$:

$$t_0^{(z)} = -t_{N_y}^{(y)}, t_1^{(z)} = -t_{N_y-1}^{(y)}, \ldots, t_{N_y}^{(z)} = -t_0^{(y)}.$$

- We then form N_y z-ranges:

$$[t_0^{(z)}, t_1^{(z)}], [t_1^{(z)}, t_2^{(z)}], \ldots, [t_{N_y-1}^{(z)}, t_{N_y}^{(z)}].$$

- Based on the intervals $\mathbf{y}_i = [\underline{y}_i, \overline{y}_i]$, we form intervals $\mathbf{z}_i = -\mathbf{y}_i = [-\overline{y}_i, -\underline{y}_i]$.
- We apply the above algorithm for computing the lower bound to compute the value $\underline{C}_{xy}(\mathbf{x}_i, -\mathbf{y}_i)$.
- Finally, we compute $\overline{C}_{xy}$ as $\overline{C}_{xy}(\mathbf{x}_i, \mathbf{y}_i) = -\underline{C}_{xy}(\mathbf{x}_i, -\mathbf{y}_i)$.

17. Toward justification of our algorithm: known facts from calculus

- 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. Let us apply these known facts to our problem

- In general, for the point $(x_1, \ldots, x_n)$ at which a function $f(x_1, \ldots, x_n)$ attains its minimum, we have:
  - if $x_i = \underline{x}_i$, then $\dfrac{\partial f}{\partial x_i} \geq 0$;
  - if $x_i = \overline{x}_i$, then $\dfrac{\partial f}{\partial x_i} \leq 0$;
  - if $\underline{x}_i < x_i < \overline{x}_i$, then $\dfrac{\partial f}{\partial x_i} = 0$.
- For covariance C_{xy}, we have $\dfrac{\partial C_{xy}}{\partial x_i} = \dfrac{1}{n} \cdot (y_i - E_y)$.
- Thus, for the point $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ at which C_{xy} attains its minimum, we have:
  - if $x_i = \underline{x}_i$, then $y_i \geq E_y$.
  - if $x_i = \overline{x}_i$, then $y_i \leq E_y$.
  - if $\underline{x}_i < x_i < \overline{x}_i$, then $y_i = E_y$.

19. Case of $\overline{y}_i < E_y$

- *Case:* $\overline{y}_i < E_y$.
- *Reminder:*
  - if $x_i = \underline{x}_i$, then $y_i \geq E_y$.
  - if $x_i = \overline{x}_i$, then $y_i \leq E_y$.
  - if $\underline{x}_i < x_i < \overline{x}_i$, then $y_i = E_y$.
- Since $\overline{y}_i < E_y$ and $y_i \leq \overline{y}_i$, we have $y_i < E_y$.
- Thus, in this case:
  - we cannot have $x_i = \underline{x}_i$, because then we would have $y_i \geq E_y$
  - we cannot have $\underline{x}_i < x_i < \overline{x}_i$, because then we would have $y_i = E_y$.
- So, if $\overline{y}_i < E_y$, the only remaining option is $x_i = \overline{x}_i$.

20. Case of $E_y < \underline{y}_i$

- *Case:* $E_y < \underline{y}_i$.
- *Reminder:*
  - if $x_i = \underline{x}_i$, then $y_i \geq E_y$.
  - if $x_i = \overline{x}_i$, then $y_i \leq E_y$.
  - if $\underline{x}_i < x_i < \overline{x}_i$, then $y_i = E_y$.
- Since $E_y < \underline{y}_i$ and $\underline{y}_i \leq y_i$, we have $E_y < y_i$.
- Thus, in this case:
  - we cannot have $x_i = \overline{x}_i$, because then we would have $y_i \leq E_y$
  - we cannot have $\underline{x}_i < x_i < \overline{x}_i$, because then we would have $y_i = E_y$.
- So, if $E_y < \underline{y}_i$, the only remaining option is $x_i = \underline{x}_i$.

21. Cases of $\overline{x}_i < E_x$ and $E_x < \underline{x}_i$

- We have shown that:
  - if $\overline{y}_i < E_y$, then $x_i = \overline{x}_i$;
  - if $E_y < \underline{y}_i$, then $x_i = \underline{x}_i$.
- We can similarly conclude that:
  - if $\overline{x}_i < E_x$, then $y_i = \overline{y}_i$;
  - if $E_x < \underline{x}_i$, then $y_i = \underline{y}_i$.
- So, we can tell exactly where the min is attained if:
  - the interval $\mathbf{x}_i$ is either completely to the left or to the right of E_x, and
  - the interval $\mathbf{y}_i$ is either completely to the left or to the right of E_y,
- E.g., if $\overline{x}_i < E_x$ ($\mathbf{x}_i$ to the left of E_x) and $E_y < \underline{y}_i$ ($\mathbf{y}_i$ to the right), then min is attained for $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$.

22. Case when one of the intervals contains E_x or E_y inside

- What if one of the intervals, e.g., $\mathbf{x}_i$, is fully to the left or fully to the right of E_x, but $\mathbf{y}_i$ contains E_y inside?
- For example, if $\overline{x}_i < E_x$, this means that $y_i = \overline{y}_i$.
- Since E_y in inside the interval $[\underline{y}_i, \overline{y}_i]$, this means that $\underline{y}_i \le E_y \le \overline{y}_i$ and thus, $E_y \le y_i$.
- If $E_y < y_i$, then, as we have shown earlier, we get $x_i = \underline{x}_i$.
- One can show that the same conclusion holds when $y_i = E_y$.
- So, in this case, we also have a single pair (x_i, y_i) where the minimum can be attained: $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$.

23. Case when $(E_x, E_y) \in c$

- Where is the point (x_i, y_i) at which the minimum is attained?
- Calculus shows that (x_i, y_i) is in the union U_1 of the following three linear segments:
  - a segment where $x_i = \underline{x}_i$ and $y_i \ge E_y$;
  - a segment where $x_i = \overline{x}_i$ and $y_i \le E_y$; and
  - a segment where $\underline{x}_i < x_i < \overline{x}_i$ and $y_i = E_y$.
- Similarly, (x_i, y_i) is in the union U_2 of the following three linear segments:
  - a segment where $y_i = \underline{y}_i$ and $x_i \ge E_x$;
  - a segment where $y_i = \overline{y}_i$ and $x_i \le E_x$; and
  - a segment where $\underline{y}_i < y_i < \overline{y}_i$ and $x_i = E_x$.
- So, $(x_i, y_i) \in U_1 \cap U_2 = \{(\underline{x}_i, \underline{y}_i), (\overline{x}_i, \overline{y}_i), (E_x, E_y)\}$.

24. Case when $(E_x, E_y) \in c$ (cont-d)

- We showed that in this case, the minimum of C_{xy} is attained at $(\underline{x}_i, \underline{y}_i)$, $(\overline{x}_i, \overline{y}_i)$, or at (E_x, E_y).
- Let us show that it cannot be attained at (E_x, E_y).
- Indeed, let us then take a small Δ and replace $x_i = E_x$ with $x_i + \Delta$ and $y_i = E_y$ with $y_i - \Delta$. Then:

$$E'_x = E_x + \frac{\Delta}{n}, \quad E'_y = E_y - \frac{\Delta}{n}, \quad C'_{xy} = C_{xy} - \frac{\Delta^2}{n} \cdot \left(1 - \frac{1}{n}\right).$$

- These equalities are easy to prove if we shift all the values of x_j by $-E_x$ and all the values of y_j by $-E_y$.
- Indeed, such a shift does not change C_{xy}.
- The new value C'_{xy} is smaller than C_{xy}, while we assumed that C_{xy} is minimal: a contradiction.
- Thus, in the case when $(E_x, E_y) \in c$, the minimum can be only attained at $(\underline{x}_i, \underline{y}_i)$ or $(\overline{x}_i, \overline{y}_i)$.

25. Proof of correctness: final step

- We know that for minimizing vector $(x_1, \ldots, x_n, y_1, \ldots, y_n)$, the pair (E_x, E_y) must be contained in one of the $N_x \cdot N_y$ cells.
- We have already shown that for each cell, if the pair (E_x, E_y) is contained in this cell, then the corresponding minimizing values x_i and y_i – at which the covariance C_{xy} attains its smallest value $\underline{C}_{xy}$ – will be as above.
- Thus, the actual minimizing value will be obtained when we analyze the corresponding cell.
- So, the desired value $\underline{C}_{xy}$ will be among the values computed by the above algorithm.
- Thus, the smallest of the computed values will be exactly $\underline{C}_{xy}$.

26. Acknowledgments

This work was supported in part:

- by the National Science Foundation grants HRD-0734825 and DUE-0926721,
- by Grant 1 T36 GM078000-01 from the National Institutes of Health,
- by Grant MSM 6198898701 from MŠMT of Czech Republic, and
- by Grant 5015 "Application of fuzzy logic with operators in the knowledge based systems" from the Science and Technology Centre in Ukraine (STCU), funded by European Union.