Estimating Third Central Moment C_3 for Privacy Case under Interval and Fuzzy Uncertainty

Ali Jalal-Kamali and Vladik Kreinovich

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

1. Need for Statistical Databases

- We want to cure diseases, we want to eliminate poverty and increase education level.
- It is not always clear what causes certain diseases, which factors affect the income and the education.
- The relation between different phenomena needs to be extracted from the empirical data.
- For this purpose, we maintain large databases.
- Data coming from census help us to understand:
  - how the parents' income level affects the children's education level, and
  - how the person education level influences his or her income level.
- Medical data help us understand role of the environment, age, gender in the spread of different diseases.

2. Need to Maintain Privacy in Statistical Databases

- We rarely know beforehand which combinations of factors are important and which are not.
- Therefore, we need to be able to test different hypotheses on the data from this database.
- Different hypotheses require different characteristics.
- So, in principle, we should allow researchers to estimate the values of all these characteristics.
- The problem is that based on these values, we can inadvertently disclose confidential information.
- If we know average blood pressure of folks below certain age, then:
  - from data for two threshold ages,
  - we can extract blood pressure of a person with a given birthday.

3. Intervals as a Way to Preserve Privacy in Statistical Databases

- One way to preserve privacy is:
  - not to store the exact data values – from which a person can be identified – in the database,
  - but rather store *ranges* (intervals).
- For example:
  - instead of recording the exact age of each patient,
  - we only record whether this age is, e.g., between 0 and 10, between 10 and 20, etc.
- In general:
  - we set threshold values $t_1, \ldots, t_K$, and
  - for each person, we store only the interval $[t_i, t_{i+1}]$ that contains the corresponding value.

4. Need to Estimate Third Central Moment C_3

- To gauge asymmetry of a probability distribution, statisticians use the third central moment.
- This a good measure of symmetry, since for symmetric distributions, this moment is equal to 0.
- Based on the sample values $x_1, \ldots, x_n$, this central moment is usually estimated as

$$C_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^3, \text{ where } E \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i.$$

- Due to privacy concerns, we only know intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ containing the values x_i.
- Thus, we need to estimate the range of possible values of C_3:

$$\mathbf{C}_3 = \{C_3(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

5. Estimating Statistical Characteristics Under Interval Uncertainty: What Is Known

- Mean $E \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$, the simplest statistical characteristic, is an increasing function of all its variables.
- So, its smallest value $\underline{E}$ is attained when each of the variables x_i attains its smallest value $\underline{x}_i$:

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

- Other statistical measures are, in general, non-monotonic.
- In general, computing the values of variance, C_3, etc., under interval uncertainty is NP-hard
- For privacy case, the range of variance, covariance, and correlation can be computed in polynomial time.

6. Computing the Minimum $\underline{C}_3$ Can Be Reduced to Computing the Maximum $\overline{C}_3$

- The function $C_3(x_1, \ldots, x_n)$ is odd, i.e., satisfies the property $C(-x_1, \ldots, -x_n) = -C(x_1, \ldots, x_n)$.
- Thus, for the intervals $-\mathbf{x}_i = \{-x_i : x_i \in [\underline{x}_i, \overline{x}_i]\} = [-\overline{x}_i, -\underline{x}_i]$, we have

$$\mathbf{C}_3(-\mathbf{x}_1, \ldots, -\mathbf{x}_n) = -\mathbf{C}_3(\mathbf{x}_1, \ldots, \mathbf{x}_n).$$

- In particular, for the upper endpoint $\overline{C}_3(-\mathbf{x}_1, \ldots, -\mathbf{x}_n)$, we get:

$$\overline{C}_3(-\mathbf{x}_1, \ldots, -\mathbf{x}_n) = -\underline{C}_3(\mathbf{x}_1, \ldots, \mathbf{x}_n).$$

- Thus, if we can compute the upper endpoint for any set of intervals, we can compute the lower endpoint as

$$\underline{C}_3(\mathbf{x}_1, \ldots, \mathbf{x}_n) = -\overline{C}_3(-\mathbf{x}_1, \ldots, -\mathbf{x}_n).$$

- Because of this possibility, in the following text, we will concentrate on computing the upper endpoint $\overline{C}_3$.

7. When a Function Attains Maximum on the Interval: Known Facts from Calculus

- A function $f(x)$ attains its maximum on $[\underline{x}, \overline{x}]$ either at one of its endpoints, or at some internal point.
- If it attains is maximum at a point $x \in (\underline{x}, \overline{x})$, then its derivative at this point is 0: $\dfrac{df}{dx} = 0$.
- If maximum is at $x = \overline{x}$, then we cannot have $\dfrac{df}{dx} < 0$: then $f(\overline{x} - \Delta x) > f(\overline{x})$.
- Thus, in this case, we must have $\dfrac{df}{dx} \geq 0$.
- Similarly, if a function $f(x)$ attains its maximum at the point $x = \underline{x}$, then we must have $\dfrac{df}{dx} \leq 0$.

8. Known Facts from Calculus: Conclusion

- Thus, for each function $f(x)$, we have three possibilities for the point x where $f(x)$ attains its maximum:
  - the first possibility is that $\underline{x} < x < \overline{x}$ and $\dfrac{df}{dx} = 0$;
  - the second possibility is that $x = \overline{x}$ and $\dfrac{df}{dx} \geq 0$;
  - the third possibility is that $x = \underline{x}$ and $\dfrac{df}{dx} \leq 0$.

9. Let Us Apply These Facts to Our Problem

- Here, $\dfrac{\partial C_3}{\partial x_i} = \dfrac{3}{n} \cdot (x_i - E)^2 - \dfrac{3}{n^2} \cdot \sum\limits_{j=1}^{n} (x_j - E)^2 =$

$$\frac{3}{n^2} \cdot ((x_i - E)^2 - \sigma^2), \text{ where } \sigma^2 \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{j=1}^{n} (x_j - E)^2.$$

- So, $\dfrac{\partial C_3}{\partial x_i} = 0$ if and only if $|x_i - E| = \sigma$, i.e., if and only if

$$x_i = E - \sigma \text{ or } x_i = E + \sigma.$$

- $\dfrac{\partial C_3}{\partial x_i} \geq 0$ if and only if $|x_i - E| \geq \sigma$, i.e., if and only if

$$x_i \leq E - \sigma \text{ or } x_i \geq E + \sigma;$$

- $\dfrac{\partial C_3}{\partial x_i} \leq 0$ if and only if $|x_i - E| \leq \sigma$, i.e., if and only if

$$E - \sigma \leq x_i \leq E + \sigma.$$

10. Analysis (cont-d)

- For each i, at a point $(x_1, \ldots, x_n)$ where C_3 attains its maximum, we get one of the three options:
  1. $\underline{x}_i < x_i < \overline{x}_i$ and either $x_i = E - \sigma$ or $x_i = E + \sigma$;
  2. $x_i = \overline{x}_i$ and either $x_i \le E - \sigma$ or $x_i \ge E + \sigma$;
  3. $x_i = \underline{x}_i$ and $E - \sigma \le x_i \le E + \sigma$.
- Let $i_\pm$ denote the number of the zone containing $E \pm \sigma$.
- Let us consider all possible locations of the interval $[t_k, t_{k+1}]$ w.r.t. $E \pm \sigma$.
- If $[t_k, t_{k+1}]$ is to the right of $E + \sigma$, then we cannot have options 1 and 3, so $x_i = \overline{x}_i$.
- Similarly, for all the intervals $[t_k, t_{k+1}]$ except for $k = i_+$, we have a single option for x_i.
- For the interval $k = i_+$, we have all three possible options for each variable x_i.

11. Towards a Feasible Algorithm: Idea

- For each k, let us denote, by n_k, the number of intervals $\mathbf{x}_i$ that coincide with $[t_k, t_{k+1}]$.
- For $k = i_+$, in principle, we have three options for each of n_k indices i, to the total of 3^{n_k} possible assignments.
- This number of assignments is non-feasibly large.
- However, since all n_k intervals are identical, what matters is how many get assigned.
- In the case of $i_- < i_+$, what matters is:
  - how many values x_i get assigned the value $x_i = \underline{x}_i$; let us denote this number by $\underline{n}$;
  - how many values x_i get assigned the value $x_i = \overline{x}_i$; let us denote this number by $\overline{n}$; and
  - how many values x_i get assigned the value $x_i = E + \sigma$; this number is equal to $n - \underline{n} - \overline{n}$.

12. Towards a Feasible Algorithm (cont-d)

- Similarly, when $i_- = i_+$, what matters is:
  - how many values x_i get assigned the value $x_i = E - \sigma$; let us denote this number by n_-;
  - how many values x_i get assigned the value $x_i = E + \sigma$; let us denote this number by n_+; and
  - how many values x_i get assigned the value $x_i = \overline{x}_i$; this number is equal to $n - n_- - n_+$.
- For each combination of such values $\underline{n}$, $\overline{n}$ (or n_-, n_+), we assign values $E - \sigma$ and/or $E + \sigma$ to some x_i.
- *Problem:* we do not know the values E and σ.

13. How to Find E and σ?

- The average of all selected values x_i should be equal to E.
- So, the sum $\sum x_i$ of all selected values x_i should be equal to $n \cdot E$.
- Thus, $n \cdot E$ is a linear combination of values $E - \sigma$, $E + \sigma$, and known values like $\underline{x}_i$ and $\overline{x}_i$.
- We can use this equality to express E as a linear function of σ: $E = E(\sigma)$.
- The average value of x_i^2 should be equal to $\sigma^2 + E^2$.
- The sum $\sum x_i^2$ of the squares of all selected values x_i should be equal to $n \cdot (E^2 + \sigma^2)$.
- Plugging in $E(\sigma)$, we get a quadratic equation in terms of σ, from which we can determine σ.
- After that, we can find $E = E(\sigma)$.

14. Resulting Algorithm

- We have thresholds $t_1, \ldots, t_K$; for each k, we know the number n_k of records of the type $[t_k, t_{k+1}]$.
- Let $t_0 \stackrel{\text{def}}{=} -\infty$ and $t_{K+1} \stackrel{\text{def}}{=} +\infty$.
- For each pair of zones $i_- < i_+$, for each $\underline{n} \geq 0$, $\overline{n} \geq 0$ s.t. $\underline{n} + \overline{n} \leq n_{i_+}$, we:
  - compute the values

$$N = n - n_{i_-} - (n_{i_+} - \underline{n} - \overline{n}),$$

$$S = \sum_{k=1}^{t_{i_- -1}} n_k \cdot t_{k+1} + \sum_{k=i_-+1}^{i_+-1} (n_k \cdot t_k) + \underline{n} \cdot t_{i_+} + \overline{n} \cdot t_{i_++1} +$$

$$\sum_{k=i_++1}^{K-1} n_k \cdot t_{k+1},$$

$$M = -n_{i_-} + (n_{i_+} - \underline{n} - \overline{n});$$

Need for Statistical . . .
Need to Maintain . . .
Intervals as a Way to . . .
Need to Estimate . . .
Estimating Statistical . . .
Resulting Algorithm
Computation Time of . . .
Need for Fuzzy . . .
From Interval to Fuzzy . . .
Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 15 of 23
Go Back
Full Screen
Close
Quit

- find σ from the quadratic equation

$$n\cdot\sigma^2+n\cdot\left(\frac{S+M\cdot\sigma}{N}\right)^2=S_2+n_{i_-}\cdot\left(\frac{S+(M-N)\cdot\sigma}{N}\right)^2+$$
$$(n_{i+}-\underline{n}-\overline{n})\cdot\left(\frac{S+(M+N)\cdot\sigma}{N}\right)^2;$$

- for each solution σ, we compute $E=\dfrac{S+M\cdot\sigma}{N}$; if $E\pm\sigma$ are in the corr. zones, we compute

$$C_3=\sum_{k=1}^{t_{i_-}-1} n_k\cdot(t_{k+1}-E)^3+n_{i_-}\cdot(-\sigma)^3+$$
$$\sum_{k=i_-+1}^{i_+-1} n_k\cdot(t_k-E)^3+\underline{n}\cdot(t_{i_+}-E)^3+$$
$$\overline{n}\cdot(t_{i_++1}-E)^3+(n_{i+}-\underline{n}-\overline{n})\cdot\sigma^3+$$
$$\sum_{k=i_++1}^{K-1} n_k\cdot(t_{k+1}-E)^3.$$

15. Algorithm (cont-d)

- For each pair $i_- = i_+$, for each pair $n_- \geq 0$, $n_+ \geq 0$ s.t. $n_- + n_+ \leq n_{i_+}$, we:
  - compute the values $N = n - n_- - n_+$,

$$S = \sum_{k=1}^{t_{i_+-1}} n_k \cdot t_{k+1} + (n - n_- - n_+) \cdot t_{i_++1} +$$

$$\sum_{k=i_++1}^{K-1} n_k \cdot t_{k+1},$$

$$M = -n_- + n_+;$$

  - find σ from the quadratic equation

$$n \cdot \sigma^2 + n \cdot \left(\frac{S + M \cdot \sigma}{N}\right)^2 = S_2 + n_- \cdot (E - \sigma)^2 +$$

$$n_+ \cdot (E + \sigma)^2;$$

- for each solution σ, if $E \pm \sigma$ are in corresponding zones, we compute

$$C_3 = \sum_{k=1}^{t_{i_+-1}} n_k \cdot (t_{k+1} - E)^3 + n_- \cdot (-\sigma)^3 + n_+ \cdot \sigma^3 +$$

$$(n - n_- - n_+) \cdot (t_{i_++1} - E)^3 + \sum_{k=i_++1}^{K-1} n_k \cdot (t_{k+1} - E)^3.$$

- We then return the largest of all computed values C_3 as the desired maximum $\overline{C}_3$.

16. Computation Time of the Proposed Algorithm

- We have K zones.
- Thus, we have K^2 pairs of zones.
- For each pair of zones, we consider pairs of natural numbers $\langle \underline{n}, \overline{n} \rangle$ whose sum does not exceed n_{i_+}.
- Since $n_{i_+} \le n$, each of these numbers $\underline{n}$, $\overline{n}$ does not exceed the total number of records n.
- There are this $\le n$ possible values of $\underline{n}$, and $\le n$ possible values of $\overline{n}$.
- Therefore, the total number of such pairs does not exceed n^2.
- For each pair, computations take time $O(K)$.
- So overall, this algorithm requires time which is quadratic in n: $O(n^2)$.

17. Need for Fuzzy Uncertainty

- So far, we assumed that we know exactly whether the age is between 0 and 10, between 10 and 20, etc.
- This makes sense if we start with an exact age and replace it with an interval to preserve privacy.
- In some practical situations, we only have an expert's impression of the age.
- An expert can say that the age is most probably between 10 and 20.
- We then have membership functions s.t. $\mu_k(x) = 1$ for $x \in [t_k, t_{k+1}]$ and $\mu_k(x) > 0$ for some $x \notin [t_k, t_{k+1}]$.
- In this case, we can apply Zadeh's extension principle get a fuzzy number corresponding to C_3.

18. From Interval to Fuzzy Uncertainty

- Zadeh's extension principle for $y = f(x_1, \ldots, x_n)$ can be described via α-cuts $\mathbf{x}_i(\alpha) = \{x_i : \mu_i(x_i) \ge \alpha\}$:

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

- Thus, estimating C_3 under fuzzy uncertainty can be reduced to several interval problems corr. to different α.
- For $\alpha < 1$, we have wider (and thus, intersecting) intervals $\mathbf{t}_k(\alpha) = [\underline{t}_k(\alpha), \overline{t}_k(\alpha)]$.
- Since these intervals intersect, each value x may be covered by several intervals of this type.
- It is reasonable to assume that at most two such intervals can contain each point x.
- In other words, while we have $\overline{t}_k(\alpha) > \underline{t}_{k+1}(\alpha)$, we should also have $\overline{t}_k(\alpha) < \underline{t}_{k+2}(\alpha)$.

19. From Interval to Fuzzy Uncertainty (cont-d)

- Now, have *two* different intervals containing $E - \sigma$ and *two* different intervals containing $E + \sigma$.
- For $E - \sigma$, this is not a serious issue, this would simply mean that for both intervals, we select $E - \sigma$.
- For $E + \sigma$, we have to select *two* pairs of natural numbers corresponding to both intervals containing $E + \sigma$.
- Selecting two pairs of numbers means selecting four natural numbers $\leq n$.
- As a result, we get an algorithm similar to the above one, but with computation time $O(n^4)$.
- This is much larger than the previous $O(n^2)$ time.

20. Acknowledgments

This work was supported in part:

- by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721,
- by Grants 1 T36 GM078000-01 and 1R43TR000173-01 from the National Institutes of Health, and
- by a grant on F-transforms from the Office of Naval Research.