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
Email: ajalalkamali@miners.utep.edu
vladik@utep.edu

***Abstract*—Some probability distributions (e.g., Gaussian) are symmetric, some (e.g., lognormal) are non-symmetric (*skewed*). How can we gauge the skeweness? For symmetric distributions, the third central moment $C_3 \stackrel{\text{def}}{=} E[(x - E(x))^3]$ is equal to 0; thus, this moment is used to characterize skewness. This moment is usually estimated, based on the observed (sample) values $x_1, \ldots, x_n$, as $C_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^3$, where $E \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$. In many practical situations, we do not know the exact values of x_i. For example, to preserve privacy, the exact values are often replaced by intervals containing these values (so that we only know whether the age is under 10, between 10 and 20, etc). Different values from these intervals lead, in general, to different values of C_3; it is desirable to find the range of all such possible values. In this paper, we propose a feasible algorithm for computing this range.**

I. Introduction

Need for statistical databases. We want to cure diseases, we want to eliminate poverty and increase education level, but 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, e.g., 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 to better understand, e.g., the role of the environment, age, gender, etc. in the spread of different diseases.

Need for maintaining privacy in statistical databases. We rarely know before hand which factors (or which combinations of factors) are important and which are not. We want to extract this information from the database. Therefore, we need to be able to test different hypotheses on the data from this database.

In order to test different hypotheses, we need to be able to compute different statistical characteristics which are needed to test a hypothesis. 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 the confidential information.

For example, a researcher may conjecture that all the patients whose blood pressure is above a certain threshold have a higher risk of heart attacks, and this researcher is looking for the value of the threshold for which the correlation between blood pressure and heart attacks is the largest. One of the natural ways for a researcher to find the best threshold is to all possible thresholds $t_1 < t_2 < \ldots < t_N$; i.e., for each of these thresholds t_i, to compute the values of different statistics based on the set S_i of all the patients whose blood pressure is greater than or equal to t_i. The more different thresholds we take, the more accurate is our determination of the optimal threshold. When the thresholds are close enough, then the difference between the sets S_i and S_{i+1} may consist of a single patient – the one whose actual blood pressure is between the two consecutive thresholds. So, by comparing the means and other statistical characteristics corresponding to the two related sets, we will be able to reconstruct all the values corresponding to this particular individual patient – and if we know all the characteristics of each person, then, by knowing one of the easy-to-obtain characteristics (e.g., exact birthdate), we would thus be able to identify all the medical characteristics of each person.

In view of the possibility of such undesirable privacy violations, it is important to make sure that privacy is protected in statistical databases.

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 some threshold values $t_1, \ldots, t_K$ and ask a person whether the actual value of the corresponding quantity is in the interval $[t_1, t_2]$, in the interval $[t_2, t_3]$, ..., or in the interval $[t_{K-1}, t_K]$.

As a result, for each quantity x and for each person i, instead of the exact value x_i of the corresponding quantity, we store an *interval* $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ that contains the actual (non-stored) value x_i. Each of these intervals coincides with one of the

given ranges

$$[t_1, t_2], [t_2, t_3], \ldots, [t_{K-1}, t_K].$$

Need to estimate third central moment C_3. To gauge asymmetry of a probability distribution, statisticians use the third central moment (see, e.g., [13]), 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

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

where

$$E = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i.$$

Estimating statistical characteristics under interval uncertainty: what is known. The general problem of estimating the range of a function under interval uncertainty is known as *interval computations*; see, e.g., [5], [9].

The need for interval computations comes beyond privacy concerns: it usually comes from the fact that in many cases, data come from measurements, and measurements are never absolutely accurate; see, e.g., [12]. In other words, the measurement result $\widetilde{x}_i$ are, in general, different from the actual (unknown) values x_i of the quantities that we are measuring. Often, the only information that we know about the measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ is the upper bound Δ_i on its absolute value: $|\Delta x_i| \le \Delta_i$. In this case, after the measurement, the only only information that we have about the actual value x_i is that this value is in the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i] = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

Thus, if we use the measured values $x_1, \ldots, x_n$ to estimate the values of some auxiliary quantity $y = f(x_1, \ldots, x_n)$, we need to know the range of possible values of y:

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

In particular, if we perform a statistical analysis of the measurement results, then, for each statistical characteristic $C(x_1, \ldots, x_n)$, we need to find its range

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

For the mean E, the situation is simple: the mean 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 $\underline{x}_i$, and its largest value $\overline{E}$ is attained when each of the variables attains its largest value $\overline{x}_i$:

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

However, other statistical measures are, in general, non-monotonic. It turns out that in general, computing the values of these characteristics under interval uncertainty is NP-hard [1], [2], [11]. This means, crudely speaking, that unless P=NP (which most computable scientists believe to be wrong), no feasible (polynomial-time) algorithm is possible that would always compute the range of the corresponding characteristic under interval uncertainty. Since variance is the second central moment, similar argument applies to third central moment too.

Estimating statistical characteristics for privacy case under interval uncertainty: what is known. For privacy case, the range of variance, covariance, and correlation can be computed in polynomial time [3], [4], [7], [8].

What we do in this paper. In this paper, we show that for privacy case, the range of third central moment C_3 can also be computed in polynomial time.

II. Analysis of the Problem

Computing the minimum $\underline{C}_3$ can be reduced to computing the maximum $\overline{C}_3$. We need to compute the range $[\underline{C}_3, \overline{C}_3]$ of the moment C_3 when each variable x_i is in the corresponding interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$. 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 $\overline{C}_3$.

When a function attains maximum on the interval: known facts from calculus. A function $f(x)$ defined on an interval $[\underline{x}, \overline{x}]$ attains its maximum on this interval either at lone of its endpoints, or in some internal point of the interval. 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 it attains its maximum at the point $x = \overline{x}$, then we cannot have $\dfrac{df}{dx} < 0$, because then, for some point $x - \Delta x \in [\underline{x}, \overline{x}]$, we would have a larger value of $f(x)$. Thus, in this case, we must have $\dfrac{df}{dx} \ge 0$.

Similarly, if a function $f(x)$ attains its maximum at the point $x = \underline{x}$, then we must have $\dfrac{df}{dx} \le 0$.

Thus, for each function $f(x)$, we have three possibilities for the value x where this function attains its maximum:

- first possibility is that $\underline{x} < x < \overline{x}$ and $\dfrac{df}{dx} = 0$;

- second possibility is that $x = \overline{x}$ and $\dfrac{df}{dx} \geq 0$;
- third possibility is that $x = \underline{x}$ and $\dfrac{df}{dx} \leq 0$.

Let us apply these known facts to our problem. For C_3 we have:

$$\frac{\partial C_3}{\partial x_i} = \frac{3}{n} \cdot (x_i - E)^2 - \frac{3}{n^2} \cdot \sum_{j=1}^{n} (x_j - E)^2 =$$

$$\frac{3}{n^2} \cdot ((x_i - E)^2 - \sigma^2),$$

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

Thus, for each i, at a point $(x_1, \ldots, x_n)$ where C_3 attains its maximum, we get one of the following three options:

1) first option is that $\underline{x}_i < x_i < \overline{x}_i$ and either $x_i = E - \sigma$ or $x_i = E + \sigma$;
2) second option is that $x_i = \overline{x}_i$ and either $x_i \leq E - \sigma$ or $x_i \geq E + \sigma$;
3) third option is that $x_i = \underline{x}_i$ and $E - \sigma \leq x_i \leq E + \sigma$.

In the privacy case, each interval $\mathbf{x}_i$ coincides with one of the intervals $[t_k, t_{k+1}]$. Let i_- denote the number of the interval that contains $E - \sigma$, and let $i_+ \geq i_-$ denote the number of the interval that contains $E + \sigma$. To apply the above conclusions, let us consider all possible locations on this interval with respect to the the interval $[t_{i_-}, t_{i_-+1}]$ that contains $E - \sigma$ and the interval $[t_{i_+}, t_{i_+ + 1}]$ that contains $E + \sigma$:

1) if the interval is completely to the right of $E + \sigma$, i.e., if $i_+ + 1 < k$ and thus, $E + \sigma < t_k$, then we cannot have the first or the third options, so we must have the second option and thus, we must have $x_i = \overline{x}_i$;
2) if the interval is completely to the left of $E - \sigma$, i.e., if $k + 1 < i_-$ and thus, $t_{k+1} < E - \sigma$, then we cannot have the first or the third options, so we must have the second option and thus, we must have $x_i = \overline{x}_i$;
3) if $i_- + 1 < k < i_+$, then we have $E - \sigma < x_i < E + \sigma$; in this case, we cannot have the first and the second option, so we must have the third option, and thus, we must have $x_i = \underline{x}_i$;
4) if $k = i_- < i_+$, i.e., if the interval $\mathbf{x}_- = [t_k, t_{k+1}]$ coincides with the interval that contains $E - \sigma$, and this interval is different from the interval that contains $E + \sigma$, then we cannot have the second option, because then we would have $x_i = \overline{x}_i = t_{k+1}$, and we know that this value is larger than $E - \sigma$; similarly, we cannot have the third option, since in this case, we would have $x_i = \underline{x}_i = t_k$ and thus, we would have $x_i < E - \sigma$; so, in this case, only the first option is possible, so we must have $x_i = E - \sigma$;
5) if $k = i_- = i_+$, then we cannot have the third option, since then we would have $x_i = \underline{x}_i = t_k < E - \sigma$; thus, we must have the first or the second option, i.e., we must have $x_i = E - \sigma$, $x_i = E + \sigma$, or $x_i = \overline{x}_i$;
6) finally, if $k = i_+ > i_-$, i.e., if the interval contains $E + \sigma$, then we must consider all three possible options: $x_i = \underline{x}_i = t_k$, $x_i = E + \sigma$, and $x_i = \overline{x}_i = t_{k+1}$.

Thus, for all the intervals $[t_k, t_{k+1}]$ except for the interval corresponding to $k = i_+$, we have a single option for x_i. For the interval $k = i_+$, we have three possible options for each variable x_i.

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, good news is that since all n_k intervals are identical, it does not matter which values x_i get assigned to different values, 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 $n - \underline{n} - \overline{n}$.

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 $\underline{n}$ and $\overline{n}$ (or n_- and n_+), we assign values $E - \sigma$ and/or $E + \sigma$ to some of the variables x_i. The problem is that we do not know the values E and σ; however, we can find them if we take into account that:

- the average of all selected values x_i should be equal to E, i.e., the sum $\sum x_i$ of all selected values x_i should be equal to $n \cdot E$;
- the average value of x_i^2 should be equal to $\sigma^2 + E^2$, i.e., the sum $\sum x_i^2$ of the squares of all selected values x_i should be equal to $n \cdot (E^2 + \sigma^2)$.

Thus, we get two equations from which we can determine both the values E and σ. The first equation equates $n \cdot E$ with a linear combination of values $E - \sigma$, $E + \sigma$, and known values like $\underline{x}_i$ and $\overline{x}_i$. Thus, this equation is a linear equation in terms of E and σ. We can use this equation to express E as a linear

function of σ. Now, the second equation becomes a quadratic equation in terms of σ, from which we can determine σ.

Towards a feasible algorithm: details. For each pair with $i_- < i_+$, once we have fixed the values $\underline{n}$ and $\overline{n}$ for which $\underline{n} + \overline{n} \le n_{i_+}$, the equation $n \cdot E = \sum x_i$ takes the form

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

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

i.e., the form

$$N\cdot E = S + M\cdot\sigma,$$

where we denoted

$$N = n - n_{i_-} - (n_{i_+} - \underline{n} - \overline{n}), \tag{1}$$

$$S = \sum_{k=1}^{t_{i_--1}} n_k\cdot t_{k+1} + \sum_{k=i_-+1}^{i_+-1}(n_k\cdot t_k) + \underline{n}\cdot t_{i_+} + \overline{n}\cdot t_{i_++1} + \sum_{k=i_++1}^{K-1} n_k\cdot t_{k+1}, \tag{2}$$

and

$$M = -n_{i_-} + (n_{i_+} - \underline{n} - \overline{n}). \tag{3}$$

Thus, we conclude that

$$E = \frac{S + M\cdot\sigma}{N}, \tag{4}$$

and therefore, that

$$E - \sigma = \frac{S + (M-N)\cdot\sigma}{N} \tag{5}$$

and

$$E + \sigma = \frac{S + (M+N)\cdot\sigma}{N}. \tag{6}$$

Similarly, for the selected values x_i, the equation

$$n\cdot(\sigma^2 + E^2) = \sum x_i^2$$

takes the form

$$n\cdot(\sigma^2+E^2) = \sum_{k=1}^{t_{i_--1}} n_k\cdot t_{k+1}^2 + n_{i_-}\cdot(E-\sigma)^2 + \sum_{k=i_-+1}^{i_+-1} n_k\cdot t_k^2 +$$

$$\underline{n}\cdot t_{i_+}^2 + \overline{n}\cdot t_{i_++1}^2 + (n_{i+} - \underline{n} - \overline{n})\cdot(E+\sigma)^2 + \sum_{k=i_++1}^{K-1} n_k\cdot t_{k+1}^2.$$

Substituting the expressions (4)–(6) into this formula, we conclude that

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

$$(n_{i+} - \underline{n} - \overline{n})\cdot\left(\frac{S + (M+N)\cdot\sigma}{N}\right)^2, \tag{7}$$

where we denoted

$$S_2 = \sum_{k=1}^{t_{i_--1}} n_k\cdot t_{k+1}^2 + \sum_{k=i_-+1}^{i_+-1} n_k\cdot t_k^2 + \underline{n}\cdot t_{i_+}^2 + \overline{n}\cdot t_{i_++1}^2 + \sum_{k=i_++1}^{K-1} n_k\cdot t_{k+1}^2. \tag{8}$$

The equation (7) is a quadratic equation in terms of σ.

Similarly, for each pair with $i_- = i_+$, once we have fixed the values n_- and n_+ for which $n_- + n_+ \le n_{i_+}$, the equation $n \cdot E = \sum x_i$ takes the form

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

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

i.e., the form

$$N\cdot E = S + M\cdot\sigma,$$

where we denoted

$$N = n - n_- - n_+, \tag{9}$$

$$S = \sum_{k=1}^{t_{i_+-1}} n_k\cdot t_{k+1} + (n - n_- - n_+)\cdot t_{i_++1} + \sum_{k=i_++1}^{K-1} n_k\cdot t_{k+1}, \tag{10}$$

and

$$M = -n_- + n_+. \tag{11}$$

Thus, we conclude that E has the form (4) and thus, $E-\sigma$ and $E+\sigma$ have the form (5) and (6). Similarly, for the selected values x_i, the equation

$$n\cdot(\sigma^2 + E^2) = \sum x_i^2$$

takes the form

$$n\cdot(\sigma^2+E^2) = \sum_{k=1}^{t_{i_+-1}} n_k\cdot t_{k+1}^2 + n_-\cdot(E-\sigma)^2 + n_+\cdot(E+\sigma)^2 +$$

$$(n - n_- - n_+)\cdot t_{i_++1}^2 + \sum_{k=i_++1}^{K-1} n_k\cdot t_{k+1}^2.$$

Substituting the expressions (4)–(6) into this formula, we conclude that

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

$$n_+\cdot(E+\sigma)^2, \tag{12}$$

where we denoted

$$S_2 = \sum_{k=1}^{t_{i_+-1}} n_k\cdot t_{k+1}^2 + (n - n_- - n_+)\cdot t_{i_++1}^2 +$$

$$\sum_{k=i_+ +1}^{K-1} n_k \cdot t_{k+1}^2. \qquad (13)$$

The equation (13) is also a quadratic equation in terms of σ.

Once the find E and σ, we can compute C_3. For $i_- < i_+$, we get

$$C_3 = \sum_{k=1}^{t_{i_- -1}} n_k \cdot (t_{k+1} - E)^3 + n_{i_-} \cdot (-\sigma)^3 + \sum_{k=i_-+1}^{i_+-1} n_k \cdot (t_k - E)^3 + \underline{n} \cdot (t_{i_+} - E)^3 + \overline{n} \cdot (t_{i_+ +1} - E)^3 + (n_{i+} - \underline{n} - \overline{n}) \cdot \sigma^3 + \sum_{k=i_+ +1}^{K-1} n_k \cdot (t_{k+1} - E)^3. \qquad (14)$$

For $i_- = i_+$, we get

$$C_3 = \sum_{k=1}^{t_{i_+ -1}} n_k \cdot (t_{k+1} - E)^3 + n_- \cdot (-\sigma)^3 + n_+ \cdot \sigma^3 + (n - n_- - n_+) \cdot (t_{i_+ +1} - E)^3 + \sum_{k=i_+ +1}^{K-1} n_k \cdot (t_{k+1} - E)^3. \qquad (15)$$

III. Resulting Algorithm

Input. We have K threshold values $t_1, \ldots, t_K$ that divide the range $[\underline{t}, \overline{t}]$ of possible values of the quantity x into $K-1$ zones

$$[t_1, t_2], \ldots, [t_{K-1}, t_K],$$

where $t_1 = \underline{t}$ and $t_K = \overline{t}$.

In the databases, we have n intervals each of which is equal to one of these zones. For each k, we have n_k intervals equal to the zone $[t_k, t_{k+1}]$; here, $\sum\limits_{k=1}^{K} n_k = n$.

The values $E - \sigma$ and $E + \sigma$ may be outside the range; to describe the possible locations of these values, we add zones $(t_0, t_1]$ with $t_0 = -\infty$ and $[t_K, t_{K+1})$ with $t_{K+1} = +\infty$.

Algorithm. Since we do not know which zone i_- contains $E - \sigma$ and which zone i_+ contains $E + \sigma$, we need to consider all possible combinations of integers $i_- \le i_+$ for which $0 \le i_-$ and $i_+ \le K + 1$.

For each pair with $i_- < i_+$, we consider all pairs of natural numbers $\underline{n}$ and $\overline{n}$ for which $\underline{n} + \overline{n} \le n_{i_+}$. For each such pair of natural numbers, we:

- compute the values (1)–(3);
- find σ from the quadratic equation (7); this quadratic equation may have zero, one or two non-negative solutions σ; for each of these solutions,
  - we compute E by using the formula (4);
  - we check whether $E - \sigma \in [t_{i_-}, t_{i_- +1}]$ and whether $E + \sigma \in [t_{i+}, t_{i_+ +1}]$;
  - if these two inclusions are satisfied, we use the formula (14) to compute C_3.

For each pair with $i_- = i_+$, we consider all pairs of natural numbers n_- and n_+ for which $n_- + n_+ \le n_{i_+}$. For each such pair of natural numbers, we:

- compute the values (9)–(11);
- find σ from the quadratic equation (12); this quadratic equation may have zero, one or two non-negative solutions σ; for each of these solutions,
  - we compute E by using the formula (4);
  - we check whether $E - \sigma \in [t_{i_+}, t_{i_+ +1}]$ and whether $E + \sigma \in [t_{i+}, t_{i_+ +1}]$;
  - if these two inclusions are satisfied, we use the formula (15) to compute C_3.

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

Computation time. For each of K^2 pairs of zones, we consider pairs of natural numbers whose sum does not exceed n_{i_+} and thus, does not exceed the total number of records 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: $O(n^2)$.

IV. From Interval to Fuzzy Uncertainty

Need for fuzzy uncertainty. In the previous text, we considered a situation in which, for each record i, we know exactly which of the intervals $[t_k, t_{k+1}]$ contains the value x_i. For example, this may mean 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 this exact age with an interval to preserve privacy. In some practical situations, however, instead of the exact age or an exact height or weight, we have an expert's impression of this characteristic. An expert can say that a patient is most probably between 10 and 20 years old, but this is not crisp information: it is possible that the actual patient is, e.g., 21 years old.

How to describe and process fuzzy uncertainty. We assume that, instead of the exact intervals $[t_k, t_{k+1}]$, we have membership functions for which $\mu_k(x) = 1$ for $x \in [t_k, t_{k+1}]$ and for which positive value extend a little bit beyond t_k and beyond t_{k+1}. In this case, we can apply Zadeh's extension principle to the formula for C_3 and get a fuzzy number corresponding to the third central moment.

It is known that Zadeh's extension principle can be described in terms of α-cuts α-cuts $\mathbf{x}_i(\alpha) = \{x_i \,|\, \mu_i(x_i) \ge \alpha\}$. It is known (see, e.g., [10]) that for any function $y = f(x_1, \ldots, x_n)$, the α-cut of y is equal to

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

In particular, this means that for the third central moment C_3, we have

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

Thus, from the computational viewpoint, the problem of estimating C_3 under fuzzy uncertainty can be reduced to several similar problems for interval uncertainty – interval problems corresponding to different values α.

In view of this reduction, in the following text, we will concentrate on estimating the correlation under interval uncertainty.

Interval computation problem corresponding to $\alpha < 1$. For $\alpha = 1$, each α-cut coincides with the original interval $\mathbf{t}_k = [\underline{t}_k, \overline{t}_k] = [t_k, t_{k+1}]$. For these intervals, $\underline{t}_k = t_k = \overline{t}_{k-1}$. For this problem, we have already described the algorithm.

For $\alpha < 1$, we have wider (and thus, intersecting) intervals $\mathbf{t}_k(\alpha) = [\underline{t}_k(\alpha), \overline{t}_k(\alpha]$ for which, in general, $\underline{t}_k(\alpha) < t_k < \overline{t}_{k-1}(\alpha)$. Since these intervals intersect, each value x may be covered by several intervals of this type. It is reasonable to assume that the uncertainty is not huge, so for each point, at most two such intervals can contain this point. In other words, while we have $\overline{t}_k(\alpha) > \underline{t}_{k+1}(\alpha)$, we should also have $\overline{t}_k(\alpha) < \underline{t}_{k+2}(\alpha)$.

The difference between this situation and the previously considered situation of non-intersecting intervals is that we can 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 $E - \sigma$. However, for $E + \sigma$, this means that we have to select not just a pair of natural numbers $\underline{n}$ and $\overline{n}$ corresponding to *one* such interval, but we need to select *two* pairs of natural numbers corresponding to both intervals containing $E + \sigma$. Selecting two pairs of numbers means selecting four natural numbers $\le n$.

As a result, we get an algorithm similar to the above one, but the computation time of this algorithm is now $O(n^4)$, which is much larger than the previous $O(n^2)$ time.

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.

REFERENCES

[1] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Computing variance for interval data is NP-hard", *ACM SIGACT News*, vol. 33, no. 2, pp. 108–118, 2002.

[2] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Exact bounds on finite populations of interval data", *Reliable Computing*, vol. 11, no. 3, pp. 207–233, 2005.

[3] A. Jalal-Kamali, "Estimating correlation under interval and fuzzy uncertainty: case of hierarchical estimation", *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2012*, Berkeley, California, August 6–8, 2012.

[4] A. Jalal-Kamali, V. Kreinovich, L. Longpre, "Estimating covariance for privacy case under interval (and fuzzy) uncertainty", *Proceedings of the 2011 World Conference on Soft Computing*, San Francisco, California, May 23–26, 2011.

[5] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics*, Springer-Verlag, London, 2001.

[6] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[7] V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos, "Interval versions of statistical techniques, with applications to environmental analysis, bioinformatics, and privacy in statistical databases", *Journal of Computational and Applied Mathematics*, vol. 199, no. 2, pp. 418–423, 2007.

[8] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos, "Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity", *Reliable Computing*, vol. 12, no. 6, pp. 471–501, 2006.

[9] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM Press, Philadelphia, Pennsylvania, 2009.

[10] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, CRC Press, Boca Raton, Florida, 2005.

[11] R. Osegueda, V. Kreinovich, L. Potluri, and R. Aló, "Non-destructive testing of aerospace structures: granularity and data mining approach", *Proceedings of the IEEE International Conference on Fuzzy Sets and Systems FUZZ-IEEE'2002*, Honolulu, Hawaii, May 12–17, 2002, vol. 1, pp. 685–689.

[12] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer Verlag, New York, 2005.

[13] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2007.

[14] G. Xiang and V. Kreinovich, "Estimating variance under interval and fuzzy uncertainty: case of hierarchical estimation", In: P. Melin, O. Castillo, L. T. Aguilar, J. Kacprzyk, and W. Pedrycz (eds.), *Foundations of Fuzzy Logic and Soft Computing*, Proceedings of the World Congress of the International Fuzzy Systems Association IFSA'2007, Cancun, Mexico, June 18–21, 2007, Springer Lecture Notes on Artificial Intelligence, 2007, Vol. 4529, pp. 3–12.