3-1-2011

Estimating Covariance for Privacy Case under Interval (and Fuzzy) Uncertainty

Ali Jalal-Kamali
University of Texas at El Paso, ajalalkamali@miners.utep.edu

Vladik Kreinovich
University of Texas at El Paso, vladik@utep.edu

Luc Longpre
University of Texas at El Paso, longpre@utep.edu

Comments:
Technical Report: UTEP-CS-11-11
Published in: Ronald R. Yager, Marek Z. Reformat, Shahnaz N. Shahbazova, and Sergei Ovchinnikov (eds.), *Proceedings of the World Conference on Soft Computing*, San Francisco, CA, May 23-26, 2011.

Estimating Covariance for Privacy Case under Interval (and Fuzzy) Uncertainty

Ali Jalal-Kamali, Vladik Kreinovich, and Luc Longpré

Abstract— **One of the main objectives of collecting data in statistical databases (medical databases, census databases) is to find important correlations between different quantities. To enable researchers to looks for such correlations, we should allow them them to ask queries testing different combinations of such quantities. However, when we receive answers to many such questions, we may inadvertently disclose information about individual patients, information that should be private.**

One way to preserve privacy in statistical databases is to store *ranges* instead of the original values. For example, instead of an exact age of a patient in a medical database, we only store the information that this age is, e.g., between 60 and 70. This idea solves the privacy problem, but it make statistical analysis more complex. Different possible values from the corresponding ranges lead, in general, to different values of the corresponding statistical characteristic; it is therefore desirable to find the range of all such values.

It is known that for mean and variance, there exist feasible algorithms for computing such ranges. In this paper, we show that similar algorithms are possible for another important statistical characteristic – covariance, whose value is important in computing correlations.

I. Introduction

Need to preserve privacy in statistical databases. In order to find relations between different quantities, we collect a large amount of data. For example, we collect a large amount of medical data – to try to find correlations between instances of a certain disease and lifestyle factors that may contribute to this disease. We collect a large amount of data in a census – to see, 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.

In some cases, we are looking for commonsense correlations – e.g., between smoking and lung diseases, obesity and diabetes, etc. However, in many cases, it is not clear which factors affect a certain disease. For example, if a rare disease appears in certain areas, it may be because of the high concentration of some chemical in these areas, but we often do not know *a priori* which chemicals to look for.

For statistical databases to be most useful for such data mining, we need to allow researchers to ask arbitrary questions. However, if we simply allow these questions, we may inadvertently disclose some information about the individuals, information which is private, and which these individuals did not want to disclose to the general public when submitting information to the secure databases.

For example, if a person has a rare disease of unknown origin, a good idea is to try all possible factors that may influence the onset of this disease: age, location, profession, etc. However, once all these factors are known, we may be able to identify this person – even when her name was not listed in the database. This disclosure may prevent potential employers from hiring her, and moreover, the very fact of such a disclosure would strongly discourage all future patients from actively participating in a similar data collections.

It is therefore desirable to make sure that privacy is preserved 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).

This makes sense from the viewpoint of a statistical database. For example, while there may be a correlation between age and certain heart diseases, this correlation is rarely of the type that a person of age 62 has a much higher probability of getting this disease than a person of age 61. Usually, it is enough to know whether a person is in his or her 60s or 70s.

And this is how data is often collected: instead of asking for an exact age, we ask a person to check whether her age is, say, in between 0 and 10, 10 and 20, etc. Similarly, instead of the exact income, we ask the person to indicate into which income bracket his or her income falls.

In general, we set some threshold values $t_0, \ldots, t_N$ and ask a person whether the actual value of the corresponding quantity is in the interval $[t_0, t_1]$, in the interval $[t_1, t_2]$, …, 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 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

$$[t_0, t_1], [t_1, t_2], \ldots, [t_{N-1}, t_N].$$

Need to estimate covariance and correlation under such interval uncertainty. As we have mentioned, one of the main objectives of collecting information into a statistical database is to find correlations between different variables.

Ali Jalal-Kamali, Vladik Kreinovich, and Luc Longpré are with the Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA (email: ajalalkamali@miners.utep.edu, {vladik,longpre}@utep.edu).

This work was supported in part by the National Science Foundation grants HRD-0734825 and DUE-0926721 and by Grant 1 T36 GM078000-01 from the National Institutes of Health.

A correlation $\rho_{x,y}$ between two quantities x and y is usually defined as

$$\rho_{x,y} = \frac{C_{x,y}}{\sigma_x \cdot \sigma_y},$$

where the covariance $C_{x,y}$ and the standard deviations $\sigma_x = \sqrt{V_x}$ and $\sigma_y = \sqrt{V_y}$ are defined as follows:

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

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

where

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

Because of the privacy concerns, we do not store the actual values x_i and y_i. Instead, we store the intervals $\mathbf{x}_i$ and $\mathbf{y}_i$. Different values of x_i and y_i from these intervals lead, in general, to different values of covariance and correlation. It is therefore desirable to find the *range* of possible values of these characteristics $C(x_1, \ldots, x_n, y_1, \ldots, y_n)$ when $x_i \in \mathbf{x}_i$ and $y_i \in \mathbf{y}_i$:

$$\mathbf{C} = \{C(x_1, \ldots, x_n, y_1, \ldots, y_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n,$$
$$y_1 \in \mathbf{y}_1, \ldots, y_n \in \mathbf{y}_n\}.$$

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., [3], [7].

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., [10]. 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_x, the situation is simple: the mean is an increasing function of all its variables. So, its smallest value $\underline{E}_x$ is attained when each of the variables x_i attains its smallest value $\underline{x}_i$, and its largest value $\overline{E}_x$ is attained when each of the variables attains its largest value $\overline{x}_i$:

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

However, variance, covariance, and correlation 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], [9]. 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.

Estimating statistical characteristics for privacy case under interval uncertainty: what is known. For privacy case, the range of variance can be computed in polynomial time [5], [6].

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

Possibility of extending our results of the fuzzy case. An alternative (and more intuitive) way to preserve privacy is not to make crisp thresholds – as we did – but to have fuzzy thresholds. In other words, instead of classifying ages into 0 to 10, 10 to 20, etc., we can classify them into very young, young, etc.

This possibility goes beyond privacy preservation – because in many case, people do not know the exact values of certain characteristics, but they can provide reasonable estimates in terms of words from natural language. For example, a person usually knows his own height and weight exactly, but not the exact height and weight of his or her neighbors. However, it is easy to tell who among the neighbors are tall, short, overweight, etc. – this additional information can supplement the information that the respondents provide about themselves.

In this case, for each i, instead of an interval $\mathbf{x}_i$, we have a fuzzy number X_i that describes the corresponding word from the natural language, with a membership function $\mu_i(x_i)$ describing a degree to which a value x_i satisfies the property X_i (e.g., is young) [4], [8]. How do we handle such fuzzy information? For each desired statistical characteristic $C(x_1, \ldots, x_n)$, Zadeh's extension principle allows us to define the fuzzy value $Y = C(X_1, \ldots, X_n)$ for fuzzy inputs $X_1, \ldots, X_n$. It is known (see, e.g., [4], [8]) that under reasonable conditions, Zadeh's extension can be naturally reformulated 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 can be obtained by solving the corresponding interval computations problem.

Thus, algorithms for computing a statistical characteristic (variance, covariance, etc.) under interval uncertainty can be used to estimate the values of the same characteristic under fuzzy uncertainty as well.

II. Analysis of the Problem

Reducing maximum to minimum. 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, and 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, $\underline{C}_{xy}(x_i, -y_i) = -\overline{C}_{xy}(x_i, y_i)$.

Thus, if we know how to compute the minimum value $\underline{C}_{xy}(x_i, y_i)$, we can then compute the maximum value $\overline{C}_{xy}(x_i, y_i)$ as

$$\overline{C}_{xy}(x_i, y_i) = -\underline{C}_{xy}(x_i, -y_i).$$

Because of this reduction, in the following text, we will concentrate on computing the minimum $\underline{C}_{xy}$. In this computation, we will use known facts from calculus.

When a function attains minimum and maximum on the interval: known facts from calculus. A function $f(x)$ defined on an interval $[\underline{x}, \overline{x}]$ attains its minimum on this interval either at lone of its endpoints, or in some internal point of the interval. If it attains is minimum at a point $x \in (a, b)$, then its derivative at this point is 0: $\dfrac{df}{dx} = 0$.

If it attains its minimum at the point $x = \underline{x}$, 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 smaller value of $f(x)$. Thus, in this case, we must have $\dfrac{df}{dx} \geq 0$.

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

For the maximum, a similar thing happens. If $f(x)$ attains is maximum at a point $x \in (a, b)$, then its derivative at this point is 0: $\dfrac{df}{dx} = 0$. If it attains its maximum at the point $x = \underline{x}$, then we must have $\dfrac{df}{dx} \leq 0$. Finally, if a function $f(x)$ attains its maximum at the point $x = \overline{x}$, then we must have $\dfrac{df}{dx} \geq 0$.

Let us apply these known facts to our problem. For covariance C,

$$\frac{\partial C}{\partial x_i} = \frac{1}{n} \cdot (y_i - E_y) \text{ and } \frac{\partial C}{\partial y_i} = \frac{1}{n} \cdot (x_i - E_x).$$

By considering the covariance as a function f x_i, for the point $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ at which C attains its minimum, we can make the following conclusions:

- 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 $y_i = E_y$.

So, if $\overline{y}_i < E_y$, this means that for the value $y_i \leq \overline{y}_i$ also satisfies the inequality $y_i < E_y$. Thus, in this case:

- we cannot have $x_i = \underline{x}_i$ — because then we would have $y_i \geq E_y$; and
- 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 for x_i is $x_i = \overline{x}_i$.

Similarly, if $E_y < \underline{y}_i$, this means that the value $y_i \geq \overline{y}_i$ also satisfies the inequality $y_i > E_y$. Thus, in this case:

- we cannot have $x_i = \overline{x}_i$ — because then we would have $y_i \leq E_y$; and
- 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 for x_i is $x_i = \underline{x}_i$.

Since the covariance is symmetric with respect to changing x and y, we can similarly conclude that:

- if $\overline{x}_i < E_x$, then $y_i = \overline{y}_i$, and
- if $E_x < \underline{x}_i$, then $y_i = \underline{y}_i$.

So, 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,

then, under these conditions, we can tell exactly where the minimum is attained.

For example, if we know:

- that $\overline{x}_i < E_x$ (i.e., that the interval $\mathbf{x}_i$ is fully to the left of E_x), and
- that $E_y < \underline{y}_i$ (i.e., that the interval $\mathbf{y}_i$ is fully to the right of E_y),

then the minimum is attained when $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$.

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 $\underline{y}_i \leq E_y \leq \overline{y}_i$ and thus, $E_y \leq y_i$. 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 $y_i = \overline{y}_i$.

The only remaining case is when:

- E_x is within the interval $\mathbf{x}_i$, and
- E_y is within the interval $\mathbf{y}_i$.

In this case, as we have mentioned, the point (x_i, y_i) where the minimum is attained belongs to the union U_1 of the following three linear segments:

- a segment where $x_i = \underline{x}_i$ and $y_i \geq E_y$;
- a segment where $x_i = \overline{x}_i$ and $y_i \leq E_y$; and
- a segment where $\underline{x}_i < x_i < \overline{x}_i$ and $y_i = E_y$.

Similarly, we can conclude that this point (x_i, y_i) belongs to the union U_2 of the following three linear segments:

- a segment where $y_i = \underline{y}_i$ and $x_i \geq E_x$;
- a segment where $y_i = \overline{y}_i$ and $x_i \leq E_x$; and
- a segment where $\underline{y}_i < y_i < \overline{y}_i$ and $x_i = E_x$.

The point (x_i, y_i) belongs to both unions, so it belongs to their intersection. One can see that this intersection consists of three points: $(\underline{x}_i, \underline{y}_i)$, $(\overline{x}_i, \overline{y}_i)$, and (E_x, E_y).

Let us prove, by contradiction, that the minimum cannot be attained for the point at which $(x_i, y_i) = (E_x, E_y)$. Indeed, let us assume that this is where the minimum is attained. Let us then take a small value Δ and replace $x_i = E_x$ with $x_i + \Delta$ and $y_i = E_y$ with $y_i - \Delta$. It is easy to show that the covariance does not change when we simply shift all the value of x_j by a constant and all the values of y_j by another constant. In particular, this is true if we shift all the value of x_j by $-E_x$ and all the values of y_j by $-E_y$, i.e., if we consider new values $x'_j = x_j - E_x$ and $y'_j = y_j - E_y$. In particular, we get $x'_i = y'_i = 0$.

For the new values, $E'_x = E'_y = 0$ and thus,

$$C_{xy} = \frac{1}{n} \cdot \sum_{j=1}^{n} x_j \cdot y_j.$$

After the change, we get the new values $x''_i = x'_i + \Delta = \Delta$ and $y''_i = y'_i - \Delta = -\Delta$. We want to see how the covariance changes, i.e., what is the value C''_{xy} of the covariance:

$$C''_{xy} = \frac{1}{n} \cdot \sum_{j=1}^{n} x''_j \cdot y''_j - E''_x \cdot E''_y.$$

Since we only changed the i-th values x_i and y_i, in the first sum, only one term changes, from $x'_i \cdot y'_i = 0$ to $x''_i \cdot y''_i = \Delta \cdot (-\Delta) = -\Delta^2$. Thus,

$$\frac{1}{n} \cdot \sum_{j=1}^{n} x''_j \cdot y''_j = \frac{1}{n} \cdot \sum_{j=1}^{n} x'_j \cdot y'_j - \frac{\Delta^2}{n} = C_{xy} - \frac{\Delta^2}{n}.$$

Similarly, the new values of E_x and E_y are:

$$E''_x = \frac{1}{n} \cdot \sum_{j=1} x''_j = \frac{1}{n} \cdot \sum_{j=1} x'_j + \frac{1}{n} \cdot \Delta = \frac{\Delta}{n};$$

$$E''_y = \frac{1}{n} \cdot \sum_{j=1} y''_j = \frac{1}{n} \cdot \sum_{j=1} y'_j - \frac{1}{n} \cdot \Delta = -\frac{\Delta}{n}.$$

Thus,

$$E''_x \cdot E''_y = \frac{\Delta}{n} \cdot \left(-\frac{\Delta}{n}\right) = \frac{\Delta^2}{n^2},$$

and so,

$$C''_{xy} = \left(C_{xy} - \frac{\Delta^2}{n}\right) + \frac{\Delta^2}{n^2} = C_{xy} - \frac{\Delta^2}{n} \cdot \left(1 - \frac{1}{n}\right).$$

This new value is smaller than C_{xy}, which contradicts to our assumption that at the original values, the covariance attains its minimum.

This contradiction proves that the minimum cannot be attained at the point (E_x, E_y), and that is therefore has to be attained at one of the two points $(\underline{x}_i, \underline{y}_i)$ and $(\overline{x}_i, \overline{y}_i)$.

Towards an algorithm. We are dealing with the privacy case. This means that each input interval $\mathbf{x}_i$ is equal to one of the x-ranges $[t_k^{(x)}, t_{k+1}^{(x)}]$ corresponding to the variable x. Let us denote the total number of such ranges by N_x.

Similarly, each input interval $\mathbf{y}_i$ is equal to one of the y-ranges $[t_\ell^{(y)}, t_{\ell+1}^{(y)}]$ corresponding to the variable y. Let us denote the total number of such ranges by N_y.

Thus, on the plane (x, y), we have $N_x \cdot N_y$ cells corresponding to different possible combinations of these ranges. For the values x_i and y_i for which the covariance attains its smallest possible value $\underline{C}_{xy}$, the corresponding means (E_x, E_y) must be located in one of these $N_x \cdot N_y$ cells.

Let us fix a cell and let us assume that the minimum is attained within this cell. Then, for each i, for the interval $\mathbf{x}_i$, there are three possible options:

- this interval may coincide with the corresponding x-range; in this case, $E_x \in \mathbf{x}_i$;
- this interval may be completely to the left of this range; in this case, $\overline{x}_i \le E_x$; and
- this interval may be completely to the right of this range; in this case, $E_x \le \underline{x}_i$.

Similarly, for the interval $\mathbf{y}_i$, there are three possible options:

- this interval may coincide with the corresponding y-range; in this case, $E_y \in \mathbf{y}_i$;
- this interval may be completely to the left of this range; in this case, $\overline{y}_i \le E_y$; and
- this interval may be completely to the right of this range; in this case, $E_y \le \underline{y}_i$.

Then, for every i for which the pair of intervals $\mathbf{x}_i$ and $\mathbf{y}_i$ is different from this cell, the above arguments enables us to uniquely determine the corresponding values x_i and y_i. For each pair for which $(\mathbf{x}_i, \mathbf{y}_i)$ coincides with this cell, we have two possible locations of the minimum: $(\underline{x}_i, \underline{y}_i)$ and $(\overline{x}_i, \overline{y}_i)$.

If we have several such intervals, then we may have arbitrary combinations of these pairs $(\underline{x}_i, \underline{y}_i)$ and $(\overline{x}_i, \overline{y}_i)$. At first glance, there are two possibilities for each i, and there can be up to n such intervals, so we can have an exponential amount 2^n of possible options.

However, the good news is that the covariance does not change if we simply reorder the intervals. Thus, if we have several intervals for which $(\mathbf{x}_i, \mathbf{y}_i)$ coincides with the given cell:

- it does not matter for which of these intervals the minimum is attained at the pair $(\underline{x}_i, \underline{y}_i)$ and for which it is attained at the pairs $(\overline{x}_i, \overline{y}_i)$;
- what matters is how many values are equal to $(\underline{x}_i, \underline{y}_i)$ (and, correspondingly, how many values are equal to $(\overline{x}_i, \overline{y}_i)$).

We can have 0, 1, ..., $\le n$ such values, so we have $\le n + 1$ such options for each cell.

So, we arrive at the following algorithm.

III. Resulting Algorithm

Input data. In the x-axis, we have $N_x + 1$ threshold values $t_0^{(x)}, t_1^{(x)}, \ldots, t_{N_x}^{(x)}$ that divide the set of possible values of the quantity x into N_x x-ranges

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

Similarly, in the y-axis, we have $N_y + 1$ threshold values $t_0^{(y)}, t_1^{(y)}, \ldots, t_{N_y}^{(y)}$ that divide the set of possible values of the quantity y into N_y y-ranges

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

We also have n data points, each of which consists of:

- an interval $\mathbf{x}_i$ that coincides with one of the x-ranges, and
- an interval $\mathbf{y}_i$ that coincides with one of the y-ranges.

Our objective: to find the endpoints $\underline{C}_{xy}$ and $\overline{C}_{xy}$ of the range

$$[\underline{C}_{xy}, \overline{C}_{xy}] = \{C(x_1, \ldots, \ldots, x_n, y_1, \ldots, y_n) :$$
$$x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n, y_1 \in \mathbf{y}_1, \ldots, y_n \in \mathbf{y}_n\},$$

where

$$C(x_1, \ldots, x_n, y_1, \ldots, y_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - E_x \cdot E_y,$$

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

Algorithm for computing $\underline{C}_{xy}$. We have N_x possible x-ranges $[t_k^{(x)}, t_{k+1}^{(x)}]$ and N_y possible y-ranges $[t_\ell^{(y)}, t_{\ell+1}^{(x)}]$. By combining an x-range and a y-range, we get $N_x \cdot N_y$ cells

$$[t_k^{(x)}, t_{k+1}^{(x)}] \times [t_\ell^{(y)}, t_{\ell+1}^{(x)}].$$

In this algorithm, we analyze these cells one by one. For each cell and for each i, we assume that the pair (E_x, E_y) corresponding to the minimizing set $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ is contained in this cell.

For each i from 1 to n, for the interval $\mathbf{x}_i$, there are three possible options:

- the interval $\mathbf{x}_i$ coincides with the x-range; we will denote this option by X^0;
- the interval $\mathbf{x}_i$ is completely to the left of the x-range; we will denote this option by X^-;
- the interval $\mathbf{x}_i$ is completely to the right of the x-range; we will denote this option by X^+.

Similarly, for the interval $\mathbf{y}_i$, there are three possible options:

- the interval $\mathbf{y}_i$ coincides with the y-range; we will denote this option by Y^0;
- the interval $\mathbf{y}_i$ is completely to the left of the y-range; we will denote this option by Y^-;
- the interval $\mathbf{y}_i$ is completely to the right of the y-range; we will denote this option by Y^+.

We thus have $3 \cdot 3 = 9$ pairs of options. For each of these pairs, we select the values x_i and y_i as follows.

Case of X^+ and Y^+. If the interval $\mathbf{x}_i$ is to the right of the x-range and the interval $\mathbf{y}_i$ is to the right of the y-range, we take:

$$x_i = \underline{x}_i \text{ and } y_i = \underline{y}_i.$$

Case of X^+ and Y^0. If the interval $\mathbf{x}_i$ is to the right of the x-range and the interval $\mathbf{y}_i$ coincides with the y-range, we take:

$$x_i = \overline{x}_i \text{ and } y_i = \underline{y}_i.$$

Case of X^+ and Y^-. If the interval $\mathbf{x}_i$ is to the right of the x-range and the interval $\mathbf{y}_i$ is to the left of the y-range, we take:

$$x_i = \overline{x}_i \text{ and } y_i = \underline{y}_i.$$

Case of X^- and Y^+. If the interval $\mathbf{x}_i$ is to the left of the x-range and the interval $\mathbf{y}_i$ is to the right of the y-range, we take:

$$x_i = \underline{x}_i \text{ and } y_i = \overline{y}_i.$$

Case of X^- and Y^0. If the interval $\mathbf{x}_i$ is to the left of the x-range and the interval $\mathbf{y}_i$ coincides with the y-range, we take:

$$x_i = \underline{x}_i \text{ and } y_i = \overline{y}_i.$$

Case of X^- and Y^-. If the interval $\mathbf{x}_i$ is to the left of the x-range and the interval $\mathbf{y}_i$ is to the left of the y-range, we take:

$$x_i = \overline{x}_i \text{ and } y_i = \overline{y}_i.$$

Case of X^0 and Y^+. If the interval $\mathbf{x}_i$ coincides with the x-range and the interval $\mathbf{y}_i$ is to the right of the y-range, we take:

$$x_i = \underline{x}_i \text{ and } y_i = \overline{y}_i.$$

Case of X^0 and Y^-. If the interval $\mathbf{x}_i$ coincides with the x-range and the interval $\mathbf{y}_i$ is to the left of the y-range, we take:

$$x_i = \overline{x}_i \text{ and } y_i = \underline{y}_i.$$

Case of X^0 and Y^0 – and the algorithm itself. Finally, we count for how many is the interval $\mathbf{x}_i$ coincides with the x-range and the interval $\mathbf{y}_i$ coincides with the y-range, and for each integer $m = 0, 1, 2, \ldots$, we assign, to m is, 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 $y_i = \overline{y}_i$.

For each of these assignments, we compute E_x and E_y. If the value E_x is in the given x-range and the value E_y is in the selected y-range, then we compute the corresponding value C_{xy}; otherwise, this assignment is dismissed.

Finally, we find the smallest of the computed values C_{xy} and return it as the desired value $\underline{C}_{xy}$.

Proof of correctness. We know that for the 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 analyzed when we analyze the corresponding cell.

So, the desired value $\underline{C}_{xy}$ will be among the values computed by the above algorithm – and thus, the smallest of the computed values will be exactly $\underline{C}_{xy}$.

Algorithm for computing $\overline{C}_{xy}$. To compute $\overline{C}_{xy}$, we can use the fact that $\overline{C}_{xy} = -\underline{C}_{xz}$, where $z = -y$. To use this fact, we form N_y threshold values for z:

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

and N_y z-ranges

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

Then, based on the intervals $\mathbf{y}_i = [\underline{y}_i, \overline{y}_i]$, we form intervals $\mathbf{z}_i = -\mathbf{y}_i = [-\overline{y}_i, -\underline{y}_i]$. After that, we apply the above algorithm to compute the value $\underline{C}_{xz}$, and then compute $\overline{C}_{xy}$ as $\overline{C}_{xy} = -\underline{C}_{xz}$.

Computation time of this algorithm. For each of $N_x \cdot N_y$ cells, we find the values x_i and y_i for each of n pairs of intervals except for those i for which $(\mathbf{x}_i, \mathbf{y}_i)$ coincides with this cell, and then compute $C_{xy} \leq n + 1$ times – depending on the number (0, 1, 2, …) of such coinciding is for which the minimum is attained at $(\underline{x}_i, \underline{y}_i)$.

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 a constant time $O(1)$, and so, for each cell, the total computation time is $O(n)$. Thus, for all $N_x \cdot N_y$ cells, we need time

$$O(N_x \cdot N_y \cdot n).$$

Discussion. Usually, the number of x-ranges and the number 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.

Thus, the above algorithm is not only feasible, it is *(asymptotically) optimal* – in the sense that it requires the smallest possible order of computation time $O(n)$.

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] 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.

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

[5] 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.

[6] 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.

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

[8] H. T. Nguyen and E. A. Walker, *First Course on Fuzzy Logic*, CRC Press, Boca Raton, Florida, 2006.

[9] R. Osegueda, V. Kreinovich, L. Potluri, and R. Al'o, "Non-destructive testing of aerospace structures: granularity and data mining approach". *Proc. FUZZ-IEEE'2002*, Honolulu, Hwaii, May 12–17, 2002, vol. 1, pp. 685–689.

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