ESTIMATING COVARIANCE UNDER INTERVAL UNCERTAINTY IN PRIVACY-PROTECTED STATISTICAL DATABASES

RAJ KIRAN KANDATHI

Computer Science Department

APPROVED:

Vladik Kreinovich, Ph.D., Chair

Luc Longpré, Ph.D.

Scott A. Starks, Ph.D.

Charles H. Ambler, Ph.D.
Dean of the Graduate School

ESTIMATING COVARIANCE UNDER INTERVAL UNCERTAINTY IN PRIVACY-PROTECTED STATISTICAL DATABASES

by

Raj Kiran Kandathi B. Tech

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Computer Science Department

THE UNIVERSITY OF TEXAS AT EL PASO

December 2004

Acknowledgements

This graduate thesis marks a great achievement of my life, which would not have been possible without the immense support and great affection of several people. I would like to extend my sincere gratitude to all of them. I am massively indebted to Dr. Vladik Kreinovich, who has given me so much guidance, insight and encouragement throughout this work. Dr. Kreinovich's inspiration, patience and great efforts to explain things beyond doubt, are certainly the main reasons that made me complete my thesis.

My special thanks to Dr. Luc Longpré and Dr. Scott Starks for serving on my committee.

I thank my parents Mr. Venu Gopal and Mrs. Sarada and my sister Swathi and my brother Bharat for their encouragement and support.

I must in particular thank my friend Ms. Kavitha without whose support and encouragement, I would not have finished this thesis.

I am also grateful to all my friends for being a constant support.

Abstract

Due to measurement uncertainty, often, instead of the actual values x_i of the measured quantities, we only know the intervals $\mathbf{x_i} = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$, where $\widetilde{x}_i$ is the measured value and Δ_i is the upper bound on the measurement error (provided, e.g., by the manufacturer of the measuring instrument). In such situations, instead of the exact value of the sample statistics such as covariance $C_{x,y}$, we can only have an interval $\mathbf{C}_{x,y}$ of possible values of this statistic. It is known that in general, computing such an interval $\mathbf{C}_{x,y}$ for $C_{x,y}$ is an NP-hard problem. Previously, an efficient algorithm was known for computing this range $\mathbf{C}_{x,y}$ for the case when the measurements are accurate enough – so that the intervals corresponding to different measurements do not intersect much. In this thesis, we provide a new efficient algorithm for computing $\mathbf{C}_{x,y}$ for the case when interval uncertainty comes from the need for privacy protection in statistical databases.

Contents

Chapter 1

Introduction

When we have n results $x_1, \ldots, x_n$ of repeated measurement of the same quantity (at different points, or at different moments of time), the traditional statistical approach usually starts with computing their sample average $E_x = (x_1 + \ldots + x_n)/n$ and their (sample) variance

$$V_x = \frac{(x_1 - E_x)^2 + \ldots + (x_n - E_x)^2}{n} \tag{1}$$

(or, equivalently, the sample standard deviation $\sigma = \sqrt{V}$). If, during each measurement i, we measure the values x_i and y_i of two different quantities x and y, then we also compute their (sample) covariance

$$C_{x,y} = \frac{(x_1 - E_x) \cdot (y_1 - E_y) + \ldots + (x_n - E_x) \cdot (y_n - E_y)}{n}; \tag{2}$$

see, e.g., [20].

Measurements are never 100% accurate, so in reality, the actual value x_i of the i-th measured quantity can differ from the measurement result $\widetilde{x}_i$. Because of these *measurement errors* $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$, the measurement result $\widetilde{x}_i$ is, in general, different from the actual value x_i of the measured quantity x_i [20].

What do we know about the errors Δx_i of direct measurements? First, the manufacturer of the measuring instrument must supply us with an upper bound Δ_i on the

measurement error. If no such upper bound is supplied, this means that no accuracy is guaranteed, and the corresponding "measuring instrument" is practically useless.

In this case, once we performed a measurement and got a measurement result $\widetilde{x}_i$, we know that the actual (unknown) value x_i of the measured quantity belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

In many practical situations, we not only know the interval $[-\Delta_i, \Delta_i]$ of possible values of the measurement error; we also know the probability of different values Δx_i within this interval. This knowledge underlies the traditional engineering approach to estimating the error of indirect measurement, in which we assume that we know the probability distributions for measurement errors Δx_i.

In practice, we can determine the desired probabilities of different values of Δx_i by comparing the results of measuring with this instrument with the results of measuring the same quantity by a standard (much more accurate) measuring instrument. Since the standard measuring instrument is much more accurate than the one used, the difference between these two measurement results is practically equal to the measurement error; thus, the empirical distribution of this difference is close to the desired probability distribution for measurement error. There are two cases, however, when this determination is not done:

- First is the case of cutting-edge measurements, e.g., measurements in fundamental science. When a Hubble telescope detects the light from a distant galaxy, there is no "standard" (much more accurate) telescope floating nearby that we can use to calibrate the Hubble: the Hubble telescope is the best we have.

- The second case is the case of measurements on the shop floor. In this case, in principle, every sensor can be thoroughly calibrated, but sensor calibration is so costly – usually costing ten times more than the sensor itself – that manufacturers rarely do it.

In both cases, we have no information about the probabilities of Δx_i; the only information we have is the upper bound on the measurement error.

In this case, after we performed a measurement and got a measurement result $\widetilde{x}_i$, the only information that we have about the actual value x_i of the measured quantity is that it belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

In biomedical systems, statistical analysis of the data often leads to improvements in medical recommendations; however, to maintain privacy, we do not want to use the exact values of the patient's parameters. Instead, for each parameter, we select fixed values, and for each patient, we only keep the corresponding range. For example, instead of keeping the exact age, we only record whether the age is between 0 and 10, 10 and 20, 20 and 30, etc. We must then perform statistical analysis based on such interval data; see, e.g., [12, 26].

In such situations, for different possible values $x_i \in \mathbf{x}_i$ and $y_i \in \mathbf{y}_i$, we get different values of E_x, E_y, V_x, and $C_{x,y}$. The question is: what are the intervals $\mathbf{E}_x$, $\mathbf{V}_x$, and $\mathbf{C}_{x,y}$ of possible values of E_x, V_x, and $C_{x,y}$?

The practical importance of this question was emphasized, e.g., in [17, 18] on the example of processing geophysical data.

This range estimation problem is a specific problem related to a combination of interval and probabilistic uncertainty. Such problems – and their potential applications – have been described, in a general context, in the monographs [13, 24]; for further developments, see, e.g., [2, 3, 5, 4, 7, 9, 14, 16, 21, 22, 25] and references therein.

In this thesis, we consider this problem for the case of covariance. For covariance, the problem of computing $\mathbf{C}_{x,y}$ is, in general, NP-hard [8]. For the case when the measurements are accurate enough – so that the intervals corresponding to different measurements do not intersect much – an efficient algorithm for computing $\mathbf{C}_{x,y}$ was presented in [1]. In this thesis, we provide a new efficient algorithm for computing $\mathbf{C}_{x,y}$ for the case when interval uncertainty comes from the need for privacy protection

in statistical databases.

The structure of this thesis is as follows. In Chapter 2, we describe what was known before about the problem of computing $\mathbf{C}_{x,y}$. Specially, we describe the computational complexity of the general problem of computing $\mathbf{C}_{x,y}$; we describe the estimates resulting from linearization, and the estimates resulting from applying straightforward interval computations to the quadratic terms in the expression for covariance; finally, we describe the algorithm for computing $\mathbf{C}_{x,y}$ for the case when intervals do not intersect much. In Chapter 3, we describe our new algorithm for computing $\mathbf{C}_{x,y}$ for privacy-related interval data.

Chapter 2

What was Known Before

2.1 Computational Complexity of the Problem: General Case

The simplest possible statistical characteristic is the mean. The arithmetic average E is a monotonically increasing function of each of its n variables $x_1, \ldots, x_n$, so its smallest possible value $\underline{E}$ is attained when each value x_i is the smallest possible ($x_i = \underline{x}_i$) and its largest possible value is attained when $x_i = \overline{x}_i$ for all i. In other words, the range $\mathbf{E}$ of E is equal to $[E(\underline{x}_1, \ldots, \overline{x}_n), E(\overline{x}_1, \ldots, \overline{x}_n)]$. In other words, $\underline{E} = \dfrac{1}{n}(\underline{x}_1 + \ldots + \underline{x}_n)$ and $\overline{E} = \dfrac{1}{n}(\overline{x}_1 + \ldots + \overline{x}_n)$.

Another widely used statistic is the variance. In contrast to the mean, the dependence of the variance V on x_i is not monotonic, so the above simple idea does not work. Rather surprisingly, it turns out that the problem of computing the exact range for the variance over interval data is, in general, NP-hard [8, 11] which means, crudely speaking, that the worst-case computation time grows exponentially with n. Moreover, if we want to compute the variance range with a given accuracy ε, the problem is still NP-hard. (For a more detailed description of NP-hardness in relation

to interval uncertainty, see, e.g., [11].)

It is also known that in general, computing covariance $\mathbf{C}_{x,y}$ is NP-hard [8]. So, crudely speaking, no feasible algorithm can compute the exact range of $\mathbf{C}_{x,y}$ for all possible data sets $\mathbf{x_1}, \ldots, \mathbf{x_n}$ and $\mathbf{y_1}, \ldots, \mathbf{y_n}$.

For $\mathbf{V}$, all known algorithms lead to an excess width. Specifically, there exist feasible algorithms for computing $\underline{V}$ (see, e.g., [8]), but in general, the problem of computing $\overline{V}$ is NP-hard [8].

It is also known that in some practically important cases, feasible algorithms for computing $\overline{V}$ are possible. One such practically useful case is when the measurement accuracy is good enough so that we can tell that the different measured values $\widetilde{x}_i$ are indeed different – e.g., the corresponding intervals $\mathbf{x}_i$ do not intersect. In this case, there exists a quadratic-time algorithm for computing $\overline{V}$; see, e.g., [8].

The only thing that we know about the general case is that in general, computing covariance $\mathbf{C}_{x,y}$ is NP-hard [19]. We also know [1], that similarly to the case of variance, it is possible to compute the interval covariance when the measurement are accurate enough to enable us to distinguish between different measurement results $(\widetilde{x}_i, \widetilde{y}_i)$.

2.2 Linearization

From the practical viewpoint, often, we may not need the exact range, we can often use approximate linearization techniques. For example, when the uncertainty comes from measurement errors Δx_i, and these errors are small, we can ignore terms that are quadratic (and of higher order) in Δx_i and get reasonable estimates for the corresponding statistical characteristics. In general, in order to estimate the range of the statistic $C(x_1, \ldots, x_n)$ on the intervals $[\underline{x}_1, \overline{x}_1]$, $\ldots$, $[\underline{x}_n, \overline{x}_n]$, we expand the function C in Taylor series at the midpoint $\widetilde{x}_i \stackrel{\text{def}}{=} (\underline{x}_i + \overline{x}_i)/2$ and keep only linear terms in

this expansion. As a result, we replace the original statistic with its linearized version $C_{\rm lin}(x_1,\ldots,x_n) = C_0 - \sum\limits_{i=1}^{n} C_i\cdot\Delta x_i$, where $C_0 \stackrel{\rm def}{=} C(\widetilde{x}_1,\ldots,\widetilde{x}_n)$, $C_i \stackrel{\rm def}{=} \dfrac{\partial C}{\partial x_i}(\widetilde{x}_1,\ldots,\widetilde{x}_n)$, and $\Delta x_i \stackrel{\rm def}{=} \widetilde{x}_i - x_i$. For each i, when $x_i \in [\underline{x}_i, \overline{x}_i]$, the difference Δx_i can take all possible values from $-\Delta_i$ to Δ_i, where $\Delta_i \stackrel{\rm def}{=} (\overline{x}_i - \underline{x}_i)/2$. Thus, in the linear approximation, we can estimate the range of the characteristic C as $[C_0-\Delta, C_0+\Delta]$, where $\Delta \stackrel{\rm def}{=} \sum\limits_{i=1}^{n} |C_i|\cdot\Delta_i$.

In particular, the covariance $C_{x,y}$ is a function of $2n$ variables $x_1,\ldots,x_n,y_1,\ldots\ y_n$. For this function,

$$\frac{\partial C}{\partial x_i} = \frac{1}{n}(y_i - \bar{y}), \qquad \frac{\partial C}{\partial y_i} = \frac{1}{n}(x_i - \bar{x}).$$

Let us denote the midpoint of an interval $[\underline{x}_i, \overline{x}_i]$ by $\widetilde{x}_i$, the midpoint of an interval $[\underline{y}_i, \overline{y}_i]$ by $\widetilde{y}_i$, and the half-widths of the corresponding intervals, by Δ_{xi} and Δ_{yi}. So, in accordance with the general formula, in the linearized case, we can estimate the range $\mathbf{C}_{x,y}$ for the covariance $C_{x,y}$ as $[C_0-\Delta, C_0+\Delta]$, where

$$C_0 = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_i - \bar{\widetilde{x}})(\widetilde{y}_i - \bar{\widetilde{y}})$$

and

$$\Delta = \sum_{i=1}^{n}\left|\frac{1}{n}\left(\widetilde{y}_i - \bar{\widetilde{y}}\right)\right|\cdot\Delta_{xi} + \sum_{i=1}^{n}\left|\frac{1}{n}\left(\widetilde{x}_i - \bar{\widetilde{x}}\right)\right|\cdot\Delta_{yi}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left|\widetilde{y}_i - \bar{\widetilde{y}}\right|\cdot\Delta_{xi} + \frac{1}{n}\sum_{i=1}^{n}\left|\widetilde{x}_i - \bar{\widetilde{x}}\right|\cdot\Delta_{yi}$$

2.3 Linearization is not Always Acceptable

In some cases, linearized estimates are not sufficient: the intervals may be wide so that quadratic terms can no longer be ignored, and/or we may be in a situation where we want to guarantee that, e.g., the variance does not exceed a certain required threshold. In such situations, we need to get the exact range – or at least an enclosure for the exact range.

2.4 Interval Computations Techniques: Brief Reminder

Historically the first method for computing the enclosure for the range is the method which is sometimes called "straightforward" interval computations. This method is based on the fact that inside the computer, every algorithm consists of elementary operations (arithmetic operations, min, max, etc.). For each elementary operation $f(a, b)$, if we know the intervals $\mathbf{a}$ and $\mathbf{b}$ for a and b, we can compute the exact range $f(\mathbf{a}, \mathbf{b})$. The corresponding formulas form the so-called *interval arithmetic*. For example,

$$[\underline{a}, \overline{a}] + [\underline{b}, \overline{b}] = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]; \quad [\underline{a}, \overline{a}] - [\underline{b}, \overline{b}] = [\underline{a} - \overline{b}, \overline{a} - \underline{b}];$$

$$[\underline{a}, \overline{a}] \cdot [\underline{b}, \overline{b}] = [\min(\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b}), \max(\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b})].$$

In straightforward interval computations, we repeat the computations forming the program f step-by-step, replacing each operation with real numbers by the corresponding operation of interval arithmetic. It is known that, as a result, we get an enclosure $\mathbf{Y} \supseteq \mathbf{y}$ for the desired range.

For the mean E, we get the exact range but for the covariance, we often get the excess width $\mathbf{Y} \neq \mathbf{y}$.

2.5 Using Interval Computations to Estimate Quadratic Terms in Covariance

Let us use straightforward interval computations to estimate the ignored quadratic terms in the formula for covariance.

In terms of $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ and $\Delta y_i \stackrel{\text{def}}{=} \widetilde{y}_i - y_i$, we have $x_i = \widetilde{x}_i - \Delta x_i$ and $y_i = \widetilde{y}_i - \Delta y_i$. Therefore, for $E_x = \dfrac{1}{n}\sum_{j=1}^{n} x_j$, and $E_y = \dfrac{1}{n}\sum_{k=1}^{n} y_k$, we have

$$E_x = \frac{1}{n}\sum_{j=1}^{n}(\widetilde{x}_j - \Delta x_i) = \frac{1}{n}\sum_{j=1}^{n}\widetilde{x}_j - \frac{1}{n}\sum_{j=1}^{n}\Delta x_j = \widetilde{E}_x - \frac{1}{n}\sum_{j=1}^{n}\Delta x_j,$$

where we denoted $\widetilde{E}_x \stackrel{\text{def}}{=} \dfrac{1}{n}\sum_{j=1}^{n}\widetilde{x}_j$, and similarly,

$$E_y = \widetilde{E}_y - \frac{1}{n}\sum_{k=1}^{n}\Delta y_k.$$

Substituting the expressions for x_i, y_i, E_x, and E_y into the formula for covariance, we conclude that

$$C_{x,y} = \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{x}_i - \Delta x_i - \widetilde{E}_x + \frac{1}{n}\sum_{j=1}^{n}\Delta x_j\right)\cdot\left(\widetilde{y}_i - \Delta y_i - \widetilde{E}_y + \frac{1}{n}\sum_{k=1}^{n}\Delta y_k\right). \quad (3)$$

After multiplying the corresponding expression, we conclude that

$$C_{x,y} = \widetilde{C}_0 + L + Q,$$

where $\widetilde{C}_0$ was defined earlier, L is the sum of all the terms that are linear in Δx_i and Δy_i, and

$$\begin{aligned} Q \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\Delta x_i\cdot\Delta y_i - \left(\frac{1}{n}\sum_{i=1}^{n}\Delta x_i\right)\cdot\left(\frac{1}{n}\sum_{k=1}^{n}\Delta y_k\right) - \left(\frac{1}{n}\sum_{j=1}^{n}\Delta x_j\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\Delta y_i\right) \\ + \left(\frac{1}{n}\sum_{j=1}^{n}\Delta x_j\right)\cdot\left(\frac{1}{n}\sum_{k=1}^{n}\Delta y_k\right) \\ = \frac{1}{n}\sum_{j=1}^{n}\Delta x_i\cdot\Delta y_i - \left(\frac{1}{n}\sum_{i=1}^{n}\Delta x_i\right)\cdot\left(\frac{1}{n}\sum_{j=1}^{n}\Delta y_j\right). \end{aligned}$$

We already know that the range of the linear part L is $[-\Delta, \Delta]$. Let us apply straightforward interval computations to estimate the range of the quadratic part Q. We know that $\Delta x_i \in [-\Delta_{xi}, \Delta_{xi}]$ and $\Delta y_i \in [-\Delta_{yi}, \Delta_{yi}]$. Thus, for every i,

$$\Delta x_i \cdot \Delta y_i \in [-\Delta_{xi}, \Delta_{xi}] \cdot [-\Delta_{yi}, \Delta_{yi}] = [-\Delta_{xi} \cdot \Delta_{yi}, \Delta_{xi} \cdot \Delta_{yi}].$$

So, we can conclude that

$$\frac{1}{n}\sum_{i=1}^{n} \Delta x_i \cdot \Delta y_i \in \left[-\frac{1}{n}\sum_{i=1}^{n} \Delta_{xi} \cdot \Delta_{yi}, \frac{1}{n}\sum_{i=1}^{n} \Delta_{xi} \cdot \Delta_{yi}\right].$$

Now, $\frac{1}{n}\sum_{i=1}^{n} \Delta x_i \in \left[-\overline{\Delta}_x, \overline{\Delta}_x\right]$, where $\overline{\Delta}_x \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \Delta x_i$, and, similarly, $\frac{1}{n}\sum_{i=1}^{n} \Delta y_j \in [-\Delta_y, \Delta_x]$, where $\overline{\Delta}_y \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \Delta y_i$. Therefore,

$$-\left(\frac{1}{n}\sum_{i=1}^{n} \Delta x_i\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n} \Delta y_i\right) \in -\left[-\overline{\Delta}_x, \overline{\Delta}_x\right] \cdot \left[-\overline{\Delta}_y, \overline{\Delta}_y\right]$$
$$= \left[-\overline{\Delta}_x \cdot \overline{\Delta}_y, \overline{\Delta}_x \cdot \overline{\Delta}_y\right].$$

Thus, $Q \in \left[-\Delta^{(2)}, \Delta^{(2)}\right]$, where $\Delta^{(2)} = \frac{1}{n}\sum_{i=1}^{n} \Delta_{xi} \cdot \Delta_{yi} + \overline{\Delta_x} \cdot \overline{\Delta_y}$. Since $C_0 + L \in [C_0 - \Delta, C_0 + \Delta]$, we can conclude that $[C_0 - \Delta - \Delta^{(2)}, C_0 + \Delta + \Delta^{(2)}]$ is a guaranteed enclosure for $C_{x,y} = C_0 + L + Q$.

2.6 Efficient Algorithms for Computing the Exact Range of Covariance for Accurate Measurements

In [1], an efficient algorithm was proposed that computes the range $\mathbf{C}_{x,y}$ for the case when the measurements are accurate enough – so that the intervals corresponding to different measurements do not intersect much.

Theorem 1. [1] *There exists a polynomial-time algorithm that, given a list of n pairwise disjoint boxes $\mathbf{x}_i \times \mathbf{y}_i$ $(1 \le i \le n)$ (i.e., in which every two boxes have an empty intersection), produces the exact range $\mathbf{C}_{x,y}$ for the covariance $C_{x,y}$.*

Theorem 2. [1] *For every integer $K > 1$, there exists a polynomial-time algorithm that, given a list of n boxes $\mathbf{x}_i \times \mathbf{y}_i$ $(1 \le i \le n)$ in which $> K$ boxes always have an empty intersection, produces the exact range $\mathbf{C}_{x,y}$ for the covariance $C_{x,y}$.*

Proof. Since $C_{x,y}$ is linear in x_i, we have $C_{-x,y} = -C_{x,y}$ hence $\mathbf{C}_{-x,y} = -\mathbf{C}_{x,y}$, so $\underline{C}_{-x,y} = -\overline{C}_{x,y}$. Because of this relation, it is sufficient to provide an algorithm for computing $\underline{C}_{x,y}$: we will then compute $\overline{C}_{x,y}$ as $-\underline{C}_{-x,y}$.

The function $C_{x,y}$ is linear in each of its variables x_i and y_i. In general, a linear function $f(x)$ attains its minimum on an interval $[\underline{x}, \overline{x}]$ at one of its endpoints: at $\underline{x}$ if f is non-decreasing ($\partial f/\partial x \ge 0$) and at $\overline{x}$ if f is non-increasing ($\partial f/\partial x \le 0$). For $C_{x,y}$, we have $\partial C_{x,y}/\partial x_i = (1/n) \cdot y_i - (1/n) \cdot E_y$, so $\partial C_{x,y}/\partial x_i \ge 0$ if and only if $y_i \ge E_y$. Thus, for each i, the values x_i^m and y_i^m at which $C_{x,y}$ attains its minimum satisfy the following four properties:

1. if $x_i^m = \underline{x}_i$, then $y_i^m \ge E_y$;
2. if $x_i^m = \overline{x}_i$, then $y_i^m \le E_y$;
3. if $y_i^m = \underline{y}_i$, then $x_i^m \ge E_x$;
4. if $y_i^m = \overline{y}_i$, then $x_i^m \le E_x$.

Let us show that if we know the vector $E \stackrel{\text{def}}{=} (E_x, E_y)$, and this vector is outside the i-th box $\mathbf{b}_i \stackrel{\text{def}}{=} \mathbf{x}_i \times \mathbf{y}_i$, then we can uniquely determine the values x_i^m and y_i^m.

Indeed, the fact that $E \notin \mathbf{b}_i$ means that either $E_x \notin \mathbf{x}_i$ or $E_y \notin \mathbf{y}_i$. Without losing generality, let us assume that $E_x \notin \mathbf{x}_i$, i.e., that either $E_x < \underline{x}_i$ or $E_x > \overline{x}_i$.

If $E_x < \underline{x}_i$, then, since $\underline{x}_i \le x_i^m$, we have $E_x < x_i^m$. Hence, according to Property 4, we cannot have $y_i^m = \overline{y}_i$. Since the minimum is always attained at one of the endpoints, we thus have $y_i^m = \underline{y}_i$. Now that we know the value of y_i^m, we can use

Properties 1 and 2:

$$\text{if } \underline{y}_i \geq E_y, \text{ then } x_i^m = \underline{x}_i; \qquad \text{if } \underline{y}_i \leq E_y, \text{ then } x_i^m = \overline{x}_i.$$

Similarly, if $E_x > \overline{x}_i$, then $y_i^m = \overline{y}_i$, and:

$$\text{if } \overline{y}_i \geq E_y, \text{ then } x_i^m = \underline{x}_i; \qquad \text{if } \overline{y}_i \leq E_y, \text{ then } x_i^m = \overline{x}_i.$$

So, to compute $\underline{C}_{x,y}$, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(2n)}$, and we sort the $2n$ values $\underline{y}_i$, $\overline{y}_i$ into a sequence $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(2n)}$. We thus get $2n \times 2n$ "zones" $\mathbf{z}_{k,l} \stackrel{\text{def}}{=} [x_{(k)}, x_{(k+1)}] \times [y_{(l)}, y_{(l+1)}]$.

We know that the average E of the actual minimum values is attained in one of these zones. If we assume that $E \in \mathbf{z}_{k,l}$, i.e., in particular, that $E_x \in [x_{(k)}, x_{(k+1)}]$, then the condition $\underline{x}_i \geq E_x$ is guaranteed to be satisfied if $\underline{x}_i \geq x_{(k+1)}$. Thus, following the above arguments, we can find the values (x_i^m, y_i^m) for all the boxes $\mathbf{b}_i$ that do not contain this zone:

Table 2.1: Minimizing values of x_i and y_i

	$\overline{y}_i \leq y_{(l)}$	$\underline{y}_i \leq y_{(l)} \leq y_{(l+1)} \leq \overline{y}_i$	$y_{(l+1)} \leq \underline{y}_i$
$\overline{x}_i \leq x_{(k)}$	$(\overline{x}_i, \overline{y}_i)$	$(\underline{x}_i, \overline{y}_i)$	$(\underline{x}_i, \overline{y}_i)$
$\underline{x}_i \leq x_{(k)} \leq x_{(k+1)} \leq \overline{x}_i$	$(\overline{x}_i, \underline{y}_i)$	?	$(\underline{x}_i, \overline{y}_i)$
$x_{(k+1)} \leq \underline{x}_i$	$(\overline{x}_i, \underline{y}_i)$	$(\overline{x}_i, \underline{y}_i)$	$(\underline{x}_i, \underline{y}_i)$

As we can see, for each of $O(n^2)$ zones $\mathbf{z}_{k,l}$, the only case when we do not know the corresponding values (x_i^m, y_i^m) is when $\mathbf{b}_i$ contains this zone. All boxes $\mathbf{b}_i$ with this property have a common intersection $\mathbf{z}_{k,l}$, thus, there can be no more than K of them. For each of these $\leq K$ boxes $\mathbf{b}_i$, we try all 4 possible combinations of endpoints as the corresponding (x_i^m, y_i^m).

Thus, for each of $O(n^2)$ zones, we must try $\leq 4^K$ possible sequences of pairs (x_i^m, y_i^m). We compute each of these n-element sequences element-by-element, so computing each sequence requires $O(n)$ computational steps.

For each of these sequences, we check whether the averages E_x and E_y are indeed within this zone, and if they are, we compute the correlation. The smallest of the resulting correlations is the desired value $\underline{C}_{x,y}$.

For each of $O(n^2)$ zones, we need $O(n)$ steps, to the total of $O(n^2) \times O(n) = O(n^3)$; computing the smallest of $O(n^2)$ values requires $O(n^2)$ more steps. Thus, our algorithm computes $\underline{C}_{x,y}$ in $O(n^3)$ steps.

2.7 Remaining Problem

With respect to privacy-protected interval valued data, we cannot apply the above algorithms because we may have many values $\mathbf{x}_i$ corresponding to the same interval; e.g. if we describe an age as $[50, 60]$ or $[60, 70]$, we may have a lot of records for which $\mathbf{x}_i = (50, 60)$, and the intersection of these intervals is not empty. Since processing privacy-protected data in statistical databases is very important, we must design algorithms for computing the range $\mathbf{C}_{x,y}$ of covariance for such data.

Chapter 3

New Algorithm

3.1 Main Idea

The algorithm from [1] cannot be directly used for intervals in privacy case, because in this case, we may have many boxes $\mathbf{b}_i = \mathbf{x}_i \times \mathbf{y}_i$ that are identical and therefore have a nonempty intersection. For example, if we analyze the dependency of salary y on age x, then we may have many pairs (x_i, y_i) for which $\mathbf{x}_i = [40, 50]$ and $\mathbf{y}_i = [50, 100]$. Since we have many $K \gg 1$, the algorithm from [1] would require 2^K steps, i.e., exponentially many steps.

Our new idea is that in this case, when we have many identical boxes $\mathbf{b}_i$ that coincide with the corresponding zone $\mathbf{z}_{k,l}$, in algorithm from [1] we have to consider different combinations of endpoints corresponding to the zones. There are exponentially many such combinations; however if, e.g., we have $\mathbf{b}_1 = \mathbf{b}_2 = \ldots = \mathbf{b}_n$, then the covariance doesn't depend on which of these boxes correspond to which boundary points (x_i, y_i), only on how many points correspond to $(\underline{x}_i, \underline{y}_i)$ and how many to $(\overline{x}_i, \overline{y}_i)$.

So, instead of enumerating all 2^n combinations, it is sufficient to test $n+1$ possible

combinations:

- a combination in which, for all boxes $\mathbf{b}_i$, we select $\underline{x}_i$ and $\underline{y}_i$;
- a combination in which we select one pair $(\overline{x}_i, \overline{y}_i)$ and $n-1$ pairs $(\underline{x}_i, \underline{y}_i)$;
- a combination in which we select two pairs $(\overline{x}_i, \overline{y}_i)$ and $n-2$ pairs $(\underline{x}_i, \underline{y}_i)$;
- ...
- a combination in which, for all boxes $\mathbf{b}_i$, we select $\overline{x}_i$ and $\overline{y}_i$.

As a result, we arrive at the following algorithm.

3.2 Algorithm

- Step 1. Sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$, and eliminate the duplicates, resulting in a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(N_x)}$, where $N_x \le 2n$. We also sort the $2n$ values $\underline{y}_i$, $\overline{y}_i$, and eliminate the duplicates, resulting in a sequence $y_{(1)} \le y_{(2)} \le \ldots \le y_{(N_y)}$, where $N_y \le 2n$.
- Step 2. We form $(N_x - 1)(N_y - 1)$ "zones" $\mathbf{z}_{1,1} \stackrel{\text{def}}{=} [x_{(1)}, x_{(2)}] \times [y_{(1)}, y_{(2)}]$, $\mathbf{z}_{1,2} \stackrel{\text{def}}{=} [x_{(1)}, x_{(2)}] \times [y_{(2)}, y_{(3)}]$, ..., $\mathbf{z}_{N_x-1,N_y-1} \stackrel{\text{def}}{=} [x_{N_x-1}, x_{N_y-1}] \times [y_{N_x}, y_{N_y}]$. In general, $\mathbf{z}_{k,l} \stackrel{\text{def}}{=} [x_{(k)}, x_{(k+1)}] \times [y_{(l)}, y_{(l+1)}]$.
- Step 3. For each zone $\mathbf{z}_{k,l}$, we assume that $E = (E_x, E_y)$ belongs to $\mathbf{z}_{k,l}$ and obtain one or several sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ that lead to minimum covariance under this assumption.
- Step 3.1. Specifically, for each i, we take:

Table 3.1: Minimizing values of x_i and y_i

	$\overline{y}_i \le y_{(l)}$	$\underline{y}_i \le y_{(l)} \le y_{(l+1)} \le \overline{y}_i$	$y_{(l+1)} \le \underline{y}_i$
$\overline{x}_i \le x_{(k)}$	$(\overline{x}_i, \overline{y}_i)$	$(\underline{x}_i, \overline{y}_i)$	$(\underline{x}_i, \overline{y}_i)$
$\underline{x}_i \le x_{(k)} \le x_{(k+1)} \le \overline{x}_i$	$(\overline{x}_i, \underline{y}_i)$	Special case	$(\underline{x}_i, \overline{y}_i)$
$x_{(k+1)} \le \underline{x}_i$	$(\overline{x}_i, \underline{y}_i)$	$(\overline{x}_i, \underline{y}_i)$	$(\underline{x}_i, \underline{y}_i)$

- Step 3.2. The special case is when the box $\mathbf{b}_i = \mathbf{x}_i \times \mathbf{y}_i$ coincides with the zone. As we go over $i = 1, \ldots, n$, we form a list of all the values i for which $\mathbf{x}_i \times \mathbf{y}_i = \mathbf{z}_{k,l}$. Let $i_1, \ldots, i_m$ be values from this list. Then, we consider $m + 1$ pairs of sequences(x, y) :

  $$(x^{(0)}, y^{(0)}), (x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

  In each sequence, values x_i, y_i corresponding to $i \neq i_j$ are given by above table and other values are defined as follows:

  - when $p \le m - j$, we take $x_{i_p}^{(j)} = \underline{x}_i$ and $y_{i_p}^{(j)} = \underline{y}_i$;
  - when $p > m - j$, we take $x_{i_p}^{(j)} = \overline{x}_i$ and $y_{i_p}^{(j)} = \overline{y}_i$.

- Step 3.3. For each of these sequences, we compute $E_x^{(j)}$ and $E_y^{(j)}$ and check whether $E_x^{(j)} \in [x_{(k)}, x_{(k+1)}]$ and $E_y^{(j)} \in [y_{(l)}, y_{(l+1)}]$.

- Step 3.4. If these two conditions are satisfied, we compute the covariance $C_{x,y}$ between the corresponding sequences $x_i^{(j)}$ and $y_i^{(j)}$.

- Step 4. The smallest of the resulting covariances is the desired value $\underline{C}_{x,y}$.

- Step 5. To compute the value of $\overline{C}_{x,y}$, we take the same intervals $\mathbf{y}_i$ as before and the intervals $-\mathbf{x}_1, \ldots, -\mathbf{x}_n$ i.e.,

  $$-\mathbf{x}_1 = [-\overline{x}_1, -\underline{x}_1], \ldots, -\mathbf{x}_n = [-\overline{x}_n, -\underline{x}_n].$$

We apply the above algorithm (Steps 1–4) to compute the minimum covariance $\underline{C}_{-x,y}$ for $-\mathbf{x}_i$ and $\mathbf{y}_i$, and than compute $\overline{C}_{x,y}$ as $-\underline{C}_{-x,y}$.

3.3 Justification

How can we prove that this algorithm is correct, i.e., that it indeed computes the range $\mathbf{C}_{x,y}$ of the covariance $C_{x,y}$?

The results presented in the previous chapter show that the previously known algorithm always computes the exact range. The only reason why we cannot simply apply this algorithm to the case when intervals come from privacy in statistical databases is that in the privacy case, the previously known algorithm requires exponential time.

In Section 3.1, we have shown that when we apply the previously known algorithm, a lot of time is wasted on re-analyzing essentially the same pairs of sequence x_i, y_i again and again – to be more precise, pairs of sequences that differ only by a joint permutation of x's and y's and for which the covariance $C_{x,y}$ is the same. When we eliminate this waste and consider every such sequence exactly once – we get exactly the above algorithm. Thus, the above algorithm is also proven (and thus, guaranteed) to always produce the exact range $\mathbf{C}_{x,y}$ for the privacy case.

To complete the description of our result, we must show that in the privacy case, the new algorithm indeed requires only polynomial time.

3.4 Computational Complexity of The New Algorithm

From the algorithm, we form $\leq 2n \times 2n$ zones. For each of $O(n^2)$ zones $\mathbf{z}_{k,l}$, we find the values (x_i, y_i) for all the boxes $\mathbf{b}_i$ that do not coincide with this zone. The only

case when we do not know the corresponding values (x_i, y_i) is when $\mathbf{b}_i$ coincides with this zone. If $m \leq n$ is the number of boxes with this property, then we consider $m + 1 \leq n + 1 = O(n)$ pairs of sequences (x_i, y_i). For each pair of sequences, the computation takes $O(n)$ steps. Therefore, the overall computation time for each zone is $O(n) \times O(n)$ steps, i.e., $O(n^2)$ steps.

For each of the $O(n^2)$ zones, we need $O(n^2)$ steps, to the total of $O(n^2) \times O(n^2) = O(n^4)$. Thus the above algorithm is a polynomial time algorithm which can be used to compute the exact bounds for the covariance in the case when intervals come from the privacy case.

3.5 Example

Let us illustrate our algorithm on a simple example. In this example we have 3 measurements of x and 3 measurements of y. The values of x-intervals are: $\mathbf{x}_1 = [1, 2]$, $\mathbf{x}_2 = [3, 5]$, $\mathbf{x}_3 = [3, 5]$. The values of y-intervals are: $\mathbf{y}_1 = [0, 1]$, $\mathbf{y}_2 = [2, 3]$, $\mathbf{y}_3 = [2, 3]$.

3.5.1 Step 1

Sorting the bounds of the above intervals in ascending order and eliminating the duplicates we get the following: $N_x = N_y = 4$,

$$x_{(1)} = 1 \leq x_{(2)} = 2 \leq x_{(3)} = 3 \leq x_{(4)} = 5;$$

$$y_{(1)} = 0 \leq y_{(2)} = 1 \leq y_{(3)} = 2 \leq y_{(4)} = 3.$$

3.5.2 Step 2

We form zones $\mathbf{z}_{k,l} = [x_{(k)}, x_{(k+1)}] \times [y_{(l)}, y_{(l+1)}]$:

$$\mathbf{z}_{1,1} = [1,2] \times [0,1], \mathbf{z}_{1,2} = [1,2] \times [1,2], \mathbf{z}_{1,3} = [1,2] \times [2,3],$$
$$\mathbf{z}_{2,1} = [2,3] \times [0,1], \mathbf{z}_{2,2} = [2,3] \times [1,2], \mathbf{z}_{2,3} = [2,3] \times [2,3],$$
$$\mathbf{z}_{3,1} = [3,5] \times [0,1], \mathbf{z}_{3,2} = [3,5] \times [1,2], \mathbf{z}_{3,3} = [3,5] \times [2,3].$$

3.5.3 Step 3

In this step, for each zone $\mathbf{z}_{k,l}$, we assume that $E = (E_x, E_y)$ belongs to $\mathbf{z}_{k,l}$ and obtain one or several sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ that lead to minimum covariance under this assumption.

Zone $\mathbf{z}_{1,1}$. Let us start with the zone $\mathbf{z}_{1,1} = [1,2] \times [0,1]$ and let us describe the sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ that lead to the minimum covariance under the assumption that $E \in \mathbf{z}_{1,1}$.

For $i = 1$, we have $\mathbf{b}_1 = [1,2] \times [0,1]$. This zone coincides with the box $\mathbf{z}_{1,1} = \mathbf{b}_1$. Therefore, according to the table described in Step 3.1, it leads to the special case that is covered by Step 3.2.

For $i = 2$, we have $\mathbf{b}_2 = [3,5] \times [2,3]$. This box satisfies the conditions $x_{(k+1)} \le \underline{x}_i$ and $y_{(l+1)} \le \underline{y}_i$. Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_2 = 3$ and $y_2 = 2$.

For $i = 3$, we have $\mathbf{b}_3 = [3,5] \times [2,3]$. This box satisfies the conditions $x_{(k+1)} \le \underline{x}_i$ and $y_{(l+1)} \le \underline{y}_i$. Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i, y_i = \underline{y}_i$ i.e., $x_3 = 3$, $y_3 = 2$.

To describe x_1 and y_1, we need to go to Step 3.2 that handles the special case, when a box $\mathbf{b}_i$ coincides with the zone. In our situation, there is only one such box, so $m = 1$ and $i_1 = 1$. According to the algorithm, we have to consider $m+1 = 2$ pairs of sequences $(x, y) : (x^{(0)}, y^{(0)})$ and $(x^{(1)}, y^{(1)})$ For $j = 0$, we get $1 = p \le m - j = 1 - 0$, so

$$x_{i_p}^{(j)} = x_1^{(0)} = 1 \text{ and } y_{i_p}^{(j)} = y_1^{(0)} = 0.$$

For $j = 1$, we get $1 = p > m - j = 1 - 1 = 0$, and thus $x_1^{(1)} = 2$ and $y_1^{(1)} = 1$. So, for the zone $\mathbf{z}_{1,1}$, we must consider two pairs of sequences:

$$x^{(0)} = (x_1^{(0)}, x_2, x_3) = (1, 3, 3); y^{(0)} = (y_1^{(0)}, y_2, y_3) = (0, 2, 2)$$

$$x^{(1)} = (x_1^{(1)}, x_2, x_3) = (2, 3, 3); y^{(1)} = (y_1^{(1)}, y_2, y_3) = (1, 2, 2).$$

On Step 3.3, we compute the values of average for these sequences. We get

$$E_x^{(0)} = \frac{x_1^{(0)} + x_2 + x_3}{3} = \frac{1 + 3 + 3}{3} = \frac{7}{3}.$$

For $y_1^{(0)}, y_2, y_3$, we get,

$$E_y^{(0)} = \frac{y_1^{(0)} + y_2 + y_3}{3} = \frac{0 + 2 + 2}{3} = \frac{4}{3}.$$

The resulting vector $E^{(0)} = (E_x^{(0)}, E_y^{(0)}) = \left(\frac{7}{3}, \frac{4}{3}\right)$ does not belong to the box $\mathbf{z}_{1,1} = [1, 2] \times [0, 1]$. So, we dismiss the pair of sequences $(x^{(0)}, y^{(0)})$. For $x^{(1)}$ and $y^{(1)}$, we get

$$E_x^{(1)} = \frac{2 + 3 + 3}{3} = \frac{8}{3},$$

$$E_y^{(1)} = \frac{1 + 2 + 2}{3} = \frac{5}{3}.$$

The resulting vector $E^{(1)} = (E_x^{(1)}, E_y^{(1)}) = \left(\frac{8}{3}, \frac{5}{3}\right)$ does not belong to the box $\mathbf{z}_{1,1} = [1, 2] \times [0, 1]$. So, we dismiss the pair of sequences $(x^{(1)}, y^{(1)})$.

Zone $\mathbf{z}_{1,2}$. Now, we consider the zone $\mathbf{z}_{1,2} = [1, 2] \times [1, 2]$.

For $i = 1$, we have $\mathbf{b}_1 = [1, 2] \times [0, 1]$. Here we have $\underline{x}_i = 1 \le x_{(k)} = 1 \le x_{(k+1)} = 2 \le \overline{x}_i = 2$, and $\overline{y}_i = 0 \le y_{(l)} = 1$ Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_1 = 2$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [3, 5] \times [2, 3]$. Here $x_{(k+1)} = 2 \le \underline{x}_i = 3$, and $y_{(l+1)} = 2 \le \underline{y}_i = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_2 = 3, y_2 = 2$.

For $i = 3$, we have $\mathbf{b}_3 = [3,5] \times [2,3]$. Here $x_{(k+1)} = 2 \leq \underline{x}_i = 3$ and $y_{(l+1)} = 2 \leq \underline{y}_i = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_3 = 3, y_3 = 2$.

So, for the zone $\mathbf{z}_{1,2}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (2,3,3); y = (y_1, y_2, y_3) = (0,2,2).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{2+3+3}{3} = \frac{8}{3}; E_y = \frac{0+2+2}{3} = \frac{4}{3}.$$

The resulting vector $E^{(0)} = \left(\frac{8}{3}, \frac{5}{3}\right)$ does not belong to the box $\mathbf{z}_{1,2} = [1,2] \times [1,2]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{1,3}$. Now, we consider the zone $\mathbf{z}_{1,3} = [1,2] \times [2,3]$.

For $i = 1$, we have $\mathbf{b}_1 = [1,2] \times [0,1]$, Here $\underline{x}_i = 1 \leq x_{(k)} = 1 \leq x_{(k+1)} = 2 \leq \overline{x}_i = 2$, and $\overline{y}_i = 1 \leq y_{(l)} = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_1 = 2$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [3,5] \times [2,3]$. Here $x_{(k+1)} = 3 \leq \underline{x}_i = 2$ and $\underline{y}_i = 2 \leq y_{(l)} = 2 \leq y_{(l+1)} = 3 \leq \overline{y}_i = 3$. Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_2 = 5$ and $y_2 = 2$.

For $i = 3$, we have $\mathbf{b}_2 = [3,5] \times [2,3]$. Here $x_{(k+1)} = 2 \leq \underline{x}_i = 3$, where $\underline{x}_i = 3, x_{(k+1)} = 2$, and $\underline{y}_i = 2 \leq y_{(l)} = 2 \leq y_{(l+1)} = 3 \leq \overline{y}_i = 3$. Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_3 = 5$ and $y_3 = 2$

So, for the zone $\mathbf{z}_{1,3}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (2,5,5); y = (y_1, y_2, y_3) = (0,2,2).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{2+5+5}{3} = 4; E_y = \frac{0+2+2}{3} = \frac{4}{3}.$$

The resulting vector $E^{(0)} = \left(4, \frac{5}{3}\right)$ does not belong to the box $\mathbf{z}_{1,3} = [1,2] \times [2,3]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{2,1}$. Next, we consider the zone $\mathbf{z}_{2,1} = [2,3] \times [0,1]$.

For $i = 1$, we have $\mathbf{b}_1 = [1,2] \times [0,1]$. Here $\overline{x}_i = 2 \leq x_{(k)} = 2$ and $\underline{y}_i = 0 \leq y_{(l)} = 0 \leq y_{(l+1)} = 1 \leq \overline{y}_i = 1$. Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_1 = 1$ and $y_1 = 1$.

For $i = 2$, we have $\mathbf{b}_2 = [3,5] \times [2,3]$. Here $x_{(k+1)} = 2 \leq \underline{x}_i = 3$ and $y_{(l+1)} = 1 \leq \underline{y}_i = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_2 = 3$ and $y_2 = 2$.

For $i = 3$, we have $\mathbf{b}_2 = [3,5] \times [2,3]$. Here $x_{(k+1)} = 3 \leq \underline{x}_i = 3$ and $y_{(l+1)} = 1 \leq \underline{y}_i = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_3 = 3$ and $y_3 = 3$.

So, for the zone $\mathbf{z}_{2,1}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (1, 3, 3); y = (y_1, y_2, y_3) = (1, 2, 3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{1+3+3}{3} = \frac{7}{3}; E_y = \frac{1+2+3}{3} = 2.$$

The resulting vector $E^{(0)} = \left(\frac{7}{3}, 2\right)$ does not belong to the box $\mathbf{z}_{2,1} = [2,3] \times [0,1]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{2,2}$. Next, we consider the zone $\mathbf{z}_{2,2} = [2,3] \times [1,2]$.

For $i = 1$, we have $\mathbf{b}_1 = [1,2] \times [0,1]$. Here $\overline{x}_i = 2 \leq x_{(k)} = 2$ and $\overline{y}_i = 1 \leq y_{(l)} = 2$ where $\overline{y}_i = 1, y_{(l)} = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_1 = 2$ and $y_1 = 1$.

For $i = 2$, we have $\mathbf{b}_2 = [3, 5] \times [2, 3]$. Here $x_{(k+1)} = 3 \leq \underline{x}_i = 3$, and $y_{(l+1)} = 2 \leq \underline{y}_i = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_2 = 3$ and $y_2 = 2$.

For $i = 3$, we have $\mathbf{b}_3 = [3, 5] \times [2, 3]$, Here $x_{(k+1)} = 3 \leq \underline{x}_i = 3$ and $y_{(l+1)} = 2 \leq \underline{y}_i = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_3 = 3$ and $y_3 = 2$.

So, for the zone $\mathbf{z}_{2,2}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (2, 3, 3); y = (y_1, y_2, y_3) = (1, 2, 2).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{2+3+3}{3} = \frac{8}{3}; E_y = \frac{1+2+2}{3} = \frac{5}{3}.$$

The resulting vector $E^{(0)} = \left(\frac{8}{3}, \frac{5}{3}\right)$ belongs to the box $\mathbf{z}_{2,2} = [2, 3] \times [1, 2]$. So we compute the correlation for the resulting single pair of sequence.

$$C_{x,y} = \frac{(x_1 - E_x^{(0)}) \cdot (y_1 - E_y^{(0)}) + (x_2 - E_x^{(0)}) \cdot (y_2 - E_y^{(0)}) + (x_3 - E_x^{(0)}) \cdot (y_3 - E_y^{(0)})}{3}$$

$$= \frac{\left(2 - \frac{8}{3}\right) \cdot \left(1 - \frac{5}{3}\right) + \left(3 - \frac{8}{3}\right) \cdot \left(2 - \frac{5}{3}\right) + \left(3 - \frac{8}{3}\right) \cdot \left(2 - \frac{5}{3}\right)}{3} = \frac{2}{9}.$$

Zone $\mathbf{z}_{2,3}$. Next, we consider the zone $\mathbf{z}_{2,3} = [2, 3] \times [2, 3]$.

For $i = 1$, we have $\mathbf{b}_1 = [1, 2] \times [0, 1]$. Here $\overline{x}_i = 2 \leq x_{(k)} = 2$, and $\overline{y}_i = 1 \leq y_{(l)} = 2$ Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_1 = 2$ and $y_1 = 1.$,

For $i = 2$, we have $\mathbf{b_2} = [3, 5] \times [2, 3]$. Here $x_{(k+1)} = 3 \leq \underline{x}_i = 3$ and $\underline{y}_i = 2 \leq y_{(l)} = 2 \leq y_{(l+1)} = 3 \leq \overline{y}_i = 3$ Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_2 = 5$ and $y_2 = 2$.

For $i = 3$, we have $\mathbf{b}_3 = [3,5] \times [2,3]$. Here $x_{(k+1)} = 3 \le \underline{x}_i = 3$ and $\underline{y}_i = 2 \le y_{(l)} = 2 \le y_{(l+1)} = 3 \le \overline{y}_i = 3$ Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$, i.e., $x_3 = 5$ and $y_3 = 2$.

So, for the zone $\mathbf{z}_{2,3}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (2, 5, 5); y = (y_1, y_2, y_3) = (1, 2, 2).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{2+5+5}{3} = 4; E_y = \frac{1+2+2}{3} = \frac{5}{3}.$$

The resulting vector $E^{(0)} = \left(\frac{7}{3}, 2\right)$ does not belong to the box $\mathbf{z}_{2,3} = [2,3] \times [2,3]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{3,1}$. Next, we consider the zone $\mathbf{z}_{3,1} = [3,5] \times [0,1]$.

For $i = 1$, we have $\mathbf{b_1} = [1,2] \times [0,1]$. Here $\overline{x}_i = 2 \le x_{(k)} = 3$, and $\underline{y}_i = 0 \le y_{(l)} = 0 \le y_{(l+1)} = 1 \le \overline{y}_i = 1$ Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_1 = 1$ and $y_1 = 1$.

For $i = 2$, we have $\mathbf{b}_2 = [3,5] \times [2,3]$. Here $\underline{x}_i = 3 \le x_{(k)} = 3 \le x_{(k+1)} = 5 \le \overline{x}_i = 5$ and $y_{(l+1)} = 1 \le \underline{y}_i = 2$ Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_2 = 3$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [3,5] \times [2,3]$. Here $\underline{x}_i = 3 \le x_{(k)} = 3 \le x_{(k+1)} = 5 \le \overline{x}_i = 5$ and $y_{(l+1)} = 1 \le \underline{y}_i = 2$ Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_3 = 3$ and $y_3 = 3$.

So, for the zone $\mathbf{z}_{3,1}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (1, 3, 3); y = (y_1, y_2, y_3) = (1, 3, 3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{1+3+3}{3} = \frac{7}{3}; E_y = \frac{1+3+3}{3} = \frac{7}{3}.$$

The resulting vector $E^{(0)} = \left(\frac{7}{3}, \frac{7}{3}\right)$ does not belong to the box $\mathbf{z}_{3,1} = [3,5] \times [0,1]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{3,2}$. Next, we consider the zone $\mathbf{z}_{3,2} = [3,5] \times [1,2]$.

For $i = 1$, we have $\mathbf{b}_1 = [1,2] \times [0,1]$. Here $\overline{x}_i = 2 \leq x_{(k)} = 3$, and $\overline{y}_i = 1 \leq y_{(l)} = 1$. Therefore, according to the table from Step 3.1, we take $x_i = \overline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_1 = 2$ and $y_1 = 1$.

For $i = 2$, we have $\mathbf{b}_2 = [3,5] \times [2,3]$, Here $\underline{x}_i = 3 \leq x_{(k)} = 3 \leq x_{(k+1)} = 5 \leq \overline{x}_i = 5$ and $y_{(l+1)} = 2 \leq \underline{y}_i = 2$. Therefore, according to the table from Step 3.1, we take $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_2 = 3$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b_3} = [3,5] \times [2,3]$. Here $\underline{x}_i = 3 \leq x_{(k)} = 3 \leq x_{(k+1)} = 5 \leq \overline{x}_i = 5$ and $y_{(l+1)} = 2 \leq \underline{y}_i = 2$. Therefore, according to the table from Step 3.1, $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_3 = 3$ and $y_3 = 3$.

So, for the zone $\mathbf{z}_{3,2}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (2,3,3); y = (y_1, y_2, y_3) = (1,3,3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{2+3+3}{3} = \frac{8}{3}; E_y = \frac{1+3+3}{3} = \frac{7}{3}.$$

The resulting vector $E^{(0)} = \left(\frac{8}{3}, \frac{7}{3}\right)$ does not belong to the box $\mathbf{z}_{3,2} = [3,5] \times [1,2]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{3,3}$. Finally, we consider the zone $\mathbf{z}_{3,3} = [3,5] \times [2,3]$.

For $i = 1$, we have $\mathbf{b}_1 = [1,2] \times [0,1]$. Here $\overline{x}_i = 2 \leq x_{(k)} = 3$, and $\overline{y}_i = 1 \leq y_{(l)} = 1$. Therefore, according to the table from Step 3.1, $x_i = \overline{x}_i$ and $y_i = \overline{y}_i$, i.e., $x_1 = 2$ and $y_1 = 1$.

For $i = 2$, we have $\mathbf{b}_2 = [3,5]\times[2,3]$. Here $\underline{x}_i = 3 \le x_{(k)} = 3 \le x_{(k+1)} = 5 \le \overline{x}_i = 5$ and $\underline{y}_i = 2 \le y_{(l)} = 2 \le y_{(l+1)} = 3 \le \overline{y}_i = 3$. Therefore, according to the table from Step 3.1, this meets the special case which will be dealt in Step 3.2.

For $i = 3$, we have $\mathbf{b_3} = [3,5]\times[2,3]$. Here $\underline{x}_i = 3 \le x_{(k)} = 3 \le x_{(k+1)} = 5 \le \overline{x}_i = 5$ and $\underline{y}_i = 2 \le y_{(l)} = 2 \le y_{(l+1)} = 3 \le \overline{y}_i = 3$. Therefore, according to the table from Step 3.1, this meets the special case which will be dealt in Step 3.2.

To describe x_2, y_2, x_3, and y_3, we need to go to Step 3.2 that handles the special case when a box $\mathbf{b}_i$ coincides with a zone. In this situation, there are two such boxes, so $m = 2, i_1 = 2$, and $i_2 = 3$. According to the algorithm, we have to consider $m + 1 = 3$ pairs of sequences $(x, y) : (x^{(0)}, y^{(0)})$, $(x^{(1)}, y^{(1)})$, and $(x^{(2)}, y^{(2)})$ For $j = 0$, we get $1 = p \le m - j = 2 - 0, 2 = p \le m - j = 2 - 0$, so

$$x_{i_p}^{(j)} = x_2^{(0)} = 3 \text{ and } y_{i_p}^{(j)} = y_2^{(0)} = 2,$$

$$x_{i_p}^{(j)} = x_3^{(0)} = 3 \text{ and } y_{i_p}^{(j)} = y_3^{(0)} = 2.$$

For $j = 1$, we get $1 = p \le m - j = 2 - 1, 2 = p > m - j = 2 - 1$, so

$$x_2^{(1)} = 3, y_2^{(1)} = 2, \text{ and } x_3^{(1)} = 5, y_3^{(1)} = 3.$$

For $j = 2$, we get $1 = p > m - j = 2 - 2, 2 = p > m - j = 2 - 2$, so

$$x_2^{(2)} = 5, y_2^{(2)} = 3 \text{ and } x_3^{(2)} = 5, y_3^{(2)} = 3.$$

So, for the zone $\mathbf{z}_{3,3}$, we must consider three pairs of sequences:

$$x^{(0)} = (x_1, x_2^{(0)}, x_3^{(0)}) = (2,3,3); y^{(0)} = (y_1, y_2^{(0)}, y_3^{(0)}) = (1,2,2)$$

$$x^{(1)} = (x_1, x_2^{(1)}, x_3^{(1)}) = (2,3,5); y^{(1)} = (y_1, y_2^{(1)}, y_3^{(1)}) = (1,2,3).$$

$$x^{(2)} = (x_1, x_2^{(2)}, x_3^{(2)}) = (2,5,5); y^{(1)} = (y_1, y_2^{(1)}, y_3^{(1)}) = (1,3,3).$$

On Step 3.3, we compute the values of average for these sequences. For $(x^{(0)}, y^{(0)})$, we get:

$$E_x^{(0)} = \frac{x_1 + x_2^{(0)} + x_3^{(0)}}{3} = \frac{2+3+3}{3} = \frac{8}{3};$$
$$E_y^{(0)} = \frac{y_1 + y_2^{(0)} + y_3^{(0)}}{3} = \frac{1+2+2}{3} = \frac{5}{3}.$$

The resulting vector $E^{(0)} = (E_x^{(0)}, E_y^{(0)}) = \left(\frac{8}{3}, \frac{5}{3}\right)$ does not belong to the box $\mathbf{z}_{3,3} = [3,5] \times [2,3]$; so, we dismiss the pair of sequences $(x^{(0)}, y^{(0)})$. For $x^{(1)}$ and $y^{(1)}$, we get

$$E_x^{(1)} = \frac{2+3+5}{3} = \frac{10}{3},$$
$$E_y^{(1)} = \frac{1+2+3}{3} = 2.$$

The resulting vector $E^{(1)} = (E_x^{(1)}, E_y^{(1)}) = \left(\frac{10}{3}, 2\right)$ belongs to the box $\mathbf{z}_{3,3} = [3,5] \times [2,3]$; so, we calculate the value of covariance, i.e.,

$$C_{x,y} = \frac{\left(2 - \frac{10}{3}\right) \cdot (1-2) + \left(3 - \frac{10}{3}\right) \cdot (2-2) + \left(5 - \frac{10}{3}\right) \cdot (3-2)}{3} = 1.$$

For $x^{(2)}$ and $y^{(2)}$, we get

$$E_x^{(2)} = \frac{2+5+5}{3} = 4,$$
$$E_y^{(2)} = \frac{1+3+3}{3} = \frac{7}{3}.$$

The resulting vector $E^{(2)} = (E_x^{(2)}, E_y^{(2)}) = \left(4, \frac{7}{3}\right)$ belongs to the box $\mathbf{z}_{3,3} = [3,5] \times [2,3]$. So, we calculate the value of covariance, and we get

$$C_{x,y} = \frac{(2-4) \cdot \left(1 - \frac{7}{3}\right) + (5-4) \cdot \left(3 - \frac{7}{3}\right) + (5-4) \cdot \left(3 - \frac{7}{3}\right)}{3} = \frac{4}{3}.$$

3.5.4 Step 4

On Step 3, we computed 3 different values of covariance $C_{x,y}$: $\frac{2}{9}$, 1, and $\frac{4}{3}$. According to Step 4, $\underline{C}_{x,y}$ is the smallest of these covariances, i.e., $\underline{C}_{x,y} = \frac{2}{9}$.

3.5.5 Step 5

Now to compute the value of $\overline{C}_{x,y}$, we take the values of $\mathbf{y}_i$ as before and $-\mathbf{x_i}$. The intervals now become $-\mathbf{x}_1 = [-2,-1]$, $-\mathbf{x}_2 = [-5,-3]$, $-\mathbf{x}_3 = [-5,-3]$ and Y-intervals, $\mathbf{y}_1 = [0,1]$, $\mathbf{y}_2 = [2,3]$, $\mathbf{y}_3 = [2,3]$.

Next we apply the algorithm to compute the minimum covariance to the intervals $-\mathbf{x_i}$ and $\mathbf{y}_i$.

3.5.6 Step 1

Sorting the bounds of the above intervals in ascending order and eliminating the duplicates we get the following: $N_x = N_y = 4$,

$$x_{(1)} = -5 \leq x_{(2)} = -3 \leq x_{(3)} = -2 \leq x_{(4)} = -1;$$

$$y_{(1)} = 0 \leq y_{(2)} = 1 \leq y_{(3)} = 2 \leq y_{(4)} = 3.$$

3.5.7 Step 2

We form zones $\mathbf{z}_{k,l} = [x_{(k)}, x_{(k+1)}] \times [y_{(l)}, y_{(l+1)}]$:

$$\mathbf{z}_{1,1} = [-5,-3] \times [0,1], \mathbf{z}_{1,2} = [-5,-3] \times [1,2], \mathbf{z}_{1,3} = [-5,-3] \times [2,3],$$

$$\mathbf{z}_{2,1} = [-3,-2] \times [0,1], \mathbf{z}_{2,2} = [-3,-2] \times [1,2], \mathbf{z}_{2,3} = [-3,-2] \times [2,3],$$

$$\mathbf{z}_{3,1} = [-2,-1] \times [0,1], \mathbf{z}_{3,2} = [-2,-1] \times [1,2], \mathbf{z}_{3,3} = [-2,-1] \times [2,3].$$

3.5.8 Step 3

In this step, for each zone $\mathbf{z}_{k,l}$, we assume that $E = (E_x, E_y)$ belongs to $\mathbf{z}_{k,l}$ and obtain one or several sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ that lead to minimum covariance under this assumption.

Zone $\mathbf{z}_{1,1}$. Let us start with the zone $\mathbf{z}_{1,1} = [-5, -3] \times [0, 1]$ and let us describe the sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ that lead to the minimum covariance under the assumption that $E \in \mathbf{z}_{1,1}$.

For $i = 1$, we have $\mathbf{b}_1 = [-2, -1] \times [0, 1]$. This box satisfies the condition $x_{(k+1)} = -5 \le \underline{x}_i = -2$ and $\underline{y}_i = 0 \le y_{(l)} = 0 \le y_{(l+1)} = 1 \le \overline{y}_i = 1$. Therefore, according to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_1 = -1$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\underline{x}_i = -5 \le x_{(k)} = -5 \le x_{(k+1)} = -3 \le \overline{x}_i = -3$ and $y_{l+1} = 1 \le \underline{y}_i = 2$, Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_2 = -5$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\underline{x}_i = -5 \le x_{(k)} = -5 \le x_{(k+1)} = -3 \le \overline{x}_i = -3$ and $y_{l+1} = 1 \le \underline{y}_i = 3$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i, y_i = \overline{y}_i$ i.e., $x_3 = -5$, $y_3 = 3$.

So, for the zone $\mathbf{z}_{1,1}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (-1, -5, -5); y = (y_1, y_2, y_3) = (1, 3, 3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{-1 - 5 - 5}{3} = -\frac{11}{3}; E_y = \frac{1 + 3 + 3}{3} = \frac{7}{3}.$$

The resulting vector $E^{(0)} = \left(-\frac{11}{3}, \frac{7}{3}\right)$ does not belong to the box $\mathbf{z}_{1,1} = [-5, -3] \times [0, 1]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{1,2}$. Now, we consider the zone $\mathbf{z}_{1,2} = [-5, -3] \times [1, 2]$.

For $i = 1$, we have $\mathbf{b}_1 = [-2, -1] \times [0, 1]$. This box satisfies the condition $x_{k+1} = -3 \le \underline{x}_i = -2$ and $\overline{y}_i = 1 \le y_{(l)} = 1$ Therefore, according to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_1 = -1$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\underline{x}_i = -5 \le x_{(k)} = -5 \le x_{(k+1)} = -3 \le \overline{x}_i = -3$ and $\overline{y}_{(l+1)} = 2 \le \underline{y}_{(i)} = 2$. Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_2 = -5$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\underline{x}_i = -5 \le x_{(k)} = -5 \le x_{(k+1)} = -3 \le \overline{x}_i = -3$ and $\overline{y}_{(l+1)} = 2 \le \underline{y}_{(i)} = 2$. Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i, y_i = \overline{y}_i$ i.e., $x_3 = -5$, $y_3 = 3$.

So, for the zone $\mathbf{z}_{1,2}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (-1, -5, -5); y = (y_1, y_2, y_3) = (0, 3, 3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{-1 - 5 - 5}{3} = -\frac{11}{3}; E_y = \frac{0 + 3 + 3}{3} = 2.$$

The resulting vector $E^{(0)} = \left(-\frac{11}{3}, 2\right)$ belongs to the box $\mathbf{z}_{1,2} = [-5, -3] \times [1, 2]$. So, we compute the covariance $C_{x,y}$ for the resulting pair of sequences. The resulting value

$$\frac{\left(-1 - \left(-\frac{11}{3}\right)\right) \cdot (0 - 2) + \left(-5 - \left(-\frac{11}{3}\right)\right) \cdot (3 - 2) + \left(-5 - \left(-\frac{11}{3}\right)\right) \cdot (3 - 2)}{3}$$

is $c_{xy} = -\frac{8}{3}$.

Zone $\mathbf{z}_{1,3}$. Now, we consider the zone $\mathbf{z}_{1,3} = [-5, -3] \times [2, 3]$.

For $i = 1$, we have $\mathbf{b}_1 = [-2, -1] \times [0, 1]$. This box satisfies the condition $x_{k+1} = -3 \le \underline{x}_i = -2$ and $\overline{y}_i = 1 \le y_{(l)} = 2$ Therefore, according to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_1 = -1$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\underline{x}_i = -5 \le x_{(k)} = -5 \le x_{(k+1)} = -3 \le \overline{x}_i = -3$ and $\underline{y}_i = 2 \le y_{(l)} = 2 \le y_{(l+1)} = 3 \le$

$\overline{y}_i = 3$. Therefore, according to the table from Step 3.1, this leads to the special case which will be dealt in Step 3.2.

For $i = 3$, we have $\mathbf{b}_3 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\underline{x}_i = -5 \leq x_{(k)} = -5 \leq x_{(k+1)} = -3 \leq \overline{x}_i = -3$ and $\underline{y}_i = 2 \leq y_{(l)} = 2 \leq y_{(l+1)} = 3 \leq \overline{y}_i = 3$. Therefore, according to the table from Step 3.1, this leads to the special case which will be dealt in Step 3.2.

To describe x_2, y_2, x_3, and y_3, we need to go to Step 3.2 that handles the special case when a box $\mathbf{b}_i$ coincides with a zone. In this situation, there are two such boxes, so $m = 2, i_1 = 2$, and $i_2 = 3$. According to the algorithm, we have to consider $m + 1 = 3$ pairs of sequences $(x, y) : (x^{(0)}, y^{(0)})$, $(x^{(1)}, y^{(1)})$, and $(x^{(2)}, y^{(2)})$ For $j = 0$, we get $1 = p \leq m - j = 2 - 0, 2 = p \leq m - j = 2 - 0$, so

$$x_{i_p}^{(j)} = x_2^{(0)} = -2 \text{ and } y_{i_p}^{(j)} = y_2^{(0)} = 0,$$

$$x_{i_p}^{(j)} = x_3^{(0)} = -2 \text{ and } y_{i_p}^{(j)} = y_3^{(0)} = 0.$$

For $j = 1$, we get $1 = p \leq m - j = 2 - 1, 2 = p > m - j = 2 - 1$, so

$$x_2^{(1)} = -2, y_2^{(1)} = 0, \text{ and } x_3^{(1)} = -1, y_3^{(1)} = 1.$$

For $j = 2$, we get $1 = p > m - j = 2 - 2, 2 = p > m - j = 2 - 2$, so

$$x_2^{(2)} = -1, y_2^{(2)} = 1 \text{ and } x_3^{(2)} = -1, y_3^{(2)} = 1.$$

So, for the zone $\mathbf{z}_{1,3}$, we must consider three pairs of sequences:

$$x^{(0)} = (x_1, x_2^{(0)}, x_3^{(0)}) = (-1, -2, -2); y^{(0)} = (y_1, y_2^{(0)}, y_3^{(0)}) = (0, 0, 0)$$

$$x^{(1)} = (x_1, x_2^{(1)}, x_3^{(1)}) = (-1, -2, -1); y^{(1)} = (y_1, y_2^{(1)}, y_3^{(1)}) = (0, 0, 1).$$

$$x^{(2)} = (x_1, x_2^{(2)}, x_3^{(2)}) = (-1, -1, -1); y^{(1)} = (y_1, y_2^{(1)}, y_3^{(1)}) = (0, 1, 1).$$

On Step 3.3, we compute the values of average for these sequences for $(x^{(0)}, y^{(0)})$. We get

$$E_x^{(0)} = \frac{x_1 + x_2^{(0)} + x_3^{(0)}}{3} = \frac{-1-2-2}{3} = -\frac{5}{3};$$
$$E_y^{(0)} = \frac{y_1 + y_2^{(0)} + y_3^{(0)}}{3} = \frac{0+0+0}{3} = 0.$$

The resulting vector $E^{(0)} = (E_x^{(0)}, E_y^{(0)}) = \left(-\frac{5}{3}, 0\right)$ does not belong to the box $\mathbf{z}_{1,3} = [-5, 3] \times [2, 3]$; so, we dismiss the pair of sequences $(x^{(0)}, y^{(0)})$. For $x^{(1)}$ and $y^{(1)}$, we get

$$E_x^{(1)} = \frac{-1-2-1}{3} = -\frac{4}{3},$$
$$E_y^{(1)} = \frac{0+0+1}{3} = \frac{1}{3}.$$

The resulting vector $E^{(1)} = (E_x^{(1)}, E_y^{(1)}) = \left(-\frac{4}{3}, \frac{1}{3}\right)$ does not belong to the box $\mathbf{z}_{1,3} = [-5, -3] \times [2, 3]$; so, we dismiss the pair of sequences $(x^{(1)}, y^{(1)})$. For $x^{(2)}$ and $y^{(2)}$, we get

$$E_x^{(2)} = \frac{-1-1-1}{3} = -1,$$
$$E_y^{(2)} = \frac{0+1+1}{3} = \frac{2}{3}.$$

The resulting vector $E^{(2)} = (E_x^{(2)}, E_y^{(2)}) = \left(-1, \frac{2}{3}\right)$ does not belong to the box $\mathbf{z}_{1,3} = [-5, -3] \times [2, 3]$; so, we dismiss the pair of sequences $(x^{(1)}, y^{(1)})$.

Zone $\mathbf{z}_{2,1}$. Now, we consider the zone $\mathbf{z}_{2,1} = [-3, -2] \times [0, 1]$.

For $i = 1$, we have $\mathbf{b}_1 = [-2, -1] \times [0, 1]$. This box satisfies the condition $x_{k+1} = -2 \leq \underline{x}_i = -2$ and $\underline{y}_i = 0 \leq y_{(l)} = 0 \leq y_{(l+1)} = 1 \leq \overline{y}_i = 1$. Therefore, according to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_1 = -1$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\overline{x}_i = -3 \leq x_k = -3$ and $y_{(l+1)} = 1 \leq \underline{y}_i = 2$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_2 = -5$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\overline{x}_i = -3 \leq x_k = -3$ and $y_{(l+1)} = 1 \leq \underline{y}_i = 2$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_3 = -5$, $y_3 = 3$.

So, for the zone $\mathbf{z}_{2,1}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (-1, -5, -5); y = (y_1, y_2, y_3) = (0, 3, 3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{-1 - 5 - 5}{3} = -\frac{11}{3}; E_y = \frac{0 + 3 + 3}{3} = 2.$$

The resulting vector $E^{(0)} = \left(-\frac{11}{3}, 2\right)$ does not belong to the box $\mathbf{z}_{2,1} = [-3, -2] \times [0, 1]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{2,2}$. Now, we consider the zone $\mathbf{z}_{2,2} = [-3, -2] \times [1, 2]$.

For $i = 1$, we have $\mathbf{b}_1 = [-2, -1] \times [0, 1]$. This box satisfies the condition $x_{(k+1)} = -2 \leq \underline{x}_i = -2$ and $\overline{y}_i = 1 \leq y_{(l)} = 1$ Therefore, according to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_1 = -1$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\overline{x}_{(i)} = -3 \leq \ x_k = -3$ and $y_{(l+1)} = 2 \leq \underline{y}_{(i)} = 2$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_2 = -5$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\overline{x}_{(i)} = -3 \leq \ x_k = -3$ and $y_{(l+1)} = 2 \leq \underline{y}_i = 2$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_3 = -5$, $y_3 = 3$.

So, for the zone $\mathbf{z}_{2,2}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (-1, -5, -5); y = (y_1, y_2, y_3) = (0, 3, 3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{-1 - 5 - 5}{3} = -\frac{11}{3}; E_y = \frac{0 + 3 + 3}{3} = 2.$$

The resulting vector $E^{(0)} = \left(-\frac{11}{3}, 2\right)$ does not belong to the box $\mathbf{z}_{2,3} = [-3,-2] \times [1,2]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{2,3}$. Now, we consider the zone $\mathbf{z}_{2,3} = [-3,-2] \times [2,3]$.

For $i = 1$, we have $\mathbf{b}_1 = [-2,-1] \times [0,1]$. This box satisfies the condition $x_{(k+1)} = -2 \leq \underline{x}_i = -2$ and $\overline{y}_i = 1 \leq y_{(l+1)} = 3$. Therefore, according to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_1 = -1$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [-5,-3] \times [2,3]$. This box satisfies the conditions $\overline{x}_i = -3 \leq x_{(k)} = -3$ and $\underline{y}_i = 2 \leq y_{(l)} = 2 \leq y_{(l+1)} = 3 \leq \overline{y}_i = 3$. Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_2 = -5$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [-5,-3] \times [2,3]$. This box satisfies the conditions $\overline{x}_i = -3 \leq x_{(k)} = -3$ and $\underline{y}_i = 2 \leq y_{(l)} = 2 \leq y_{(l+1)} = 3 \leq \overline{y}_i = 3$. Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_3 = -5$, $y_3 = 3$.

So, for the zone $\mathbf{z}_{2,3}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (-1,-5,-5); y = (y_1, y_2, y_3) = (0,3,3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{-1-5-5}{3} = -\frac{11}{3}; E_y = \frac{0+3+3}{3} = 2.$$

The resulting vector $E^{(0)} = \left(-\frac{11}{3}, 2\right)$ does not belong to the box $\mathbf{z}_{2,3} = [-3,-2] \times [2,3]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{3,1}$. Now, we consider the zone $\mathbf{z}_{3,1} = [-2,-1] \times [0,1]$.

For $i = 1$, we have $\mathbf{b}_1 = [-2,-1] \times [0,1]$. This box satisfies the conditions $\underline{x}_i = -2 \leq x_{(k)} = -2 \leq x_{(k+1)} = -1 \leq \overline{x}_i = -1$ and $\underline{y}_i = 0 \leq y_{(l)} = 0 \leq y_{(l+1)} = 1 \leq \overline{y}_i = 1$. Therefore, according to the table from Step 3.1, this leads to the special case which will be dealt in Step 3.2.

For $i = 2$, we have $\mathbf{b}_2 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\overline{x}_i = -3 \leq x_{(k)} = -2$ and $y_{(l+1)} = 1 \leq \underline{y}_{(i)} = 2$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_2 = -5$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\overline{x}_i = -3 \leq x_{(k)} = -2$ and $y_{(l+1)} = 1 \leq \underline{y}_{(i)} = 2$. Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_3 = -5$, $y_3 = 3$.

To describe x_1 and y_1, we need to go to Step 3.2 that handles the special case, when a box $\mathbf{b}_i$ coincides with the zone. In our situation, there is only one such box, so $m = 1$ and $i_1 = 1$. According to the algorithm, we have to consider $m+1 = 2$ pairs of sequences $(x, y) : (x^{(0)}, y^{(0)})$ and $(x^{(1)}, y^{(1)})$ For $j = 0$, we get $1 = p \leq m - j = 1 - 0$, so

$$x_{i_p}^{(j)} = x_1^{(0)} = -2 \text{ and } y_{i_p}^{(j)} = y_1^{(0)} = 0.$$

For $j = 1$, we get $1 = p > m - j = 1 - 1 = 0$, and thus $x_1^{(1)} = -1$ and $y_1^{(1)} = 1$. So, for the zone $\mathbf{z}_{3,1}$, we must consider two pairs of sequences:

$$x^{(0)} = (x_1^{(0)}, x_2, x_3) = (-5, -5, -2); y^{(0)} = (y_1^{(0)}, y_2, y_3) = (3, 3, 0)$$

$$x^{(1)} = (x_1^{(1)}, x_2, x_3) = (-5, -5, -1); y^{(1)} = (y_1^{(1)}, y_2, y_3) = (3, 3, 1).$$

On Step 3.3, we compute the values of average for these sequences. We get

$$E_x^{(0)} = \frac{x_1^{(0)} + x_2 + x_3}{3} = \frac{-5 - 5 - 2}{3} = -4.$$

For $y_1^{(0)}, y_2, y_3$, we get,

$$E_y^{(0)} = \frac{y_1^{(0)} + y_2 + y_3}{3} = \frac{3 + 3 + 0}{3} = 2.$$

The resulting vector $E^{(0)} = (E_x^{(0)}, E_y^{(0)}) = (-4, 2)$ does not belong to the box $\mathbf{z}_{3,1} = [-2, -1] \times [0, 1]$. So, we dismiss the pair of sequences $(x^{(0)}, y^{(0)})$. For $x^{(1)}$ and $y^{(1)}$, we get

$$E_x^{(1)} = \frac{-5 - 5 - 1}{3} = -\frac{11}{3},$$

$$E_y^{(1)} = \frac{3+3+1}{3} = \frac{7}{3}.$$

The resulting vector $E^{(1)} = (E_x^{(1)}, E_y^{(1)}) = \left(-\frac{11}{3}, \frac{7}{3}\right)$ does not belong to the box $\mathbf{z}_{3,1} = [-2,-1] \times [0,1]$. So, we dismiss the pair of sequences $(x^{(1)}, y^{(1)})$.

Zone $\mathbf{z}_{3,2}$. Now, we consider the zone $\mathbf{z}_{3,2} = [-2,-1] \times [1,2]$.

For $i = 1$, we have $\mathbf{b}_1 = [-2,-1] \times [0,1]$. This box satisfies the condition $\underline{x}_i = -2 \le x_{(k)} = -2 \le x_{(k+1)} = -1 \le \overline{x}_i = -1$ and $\overline{y}_i = 1 \le y_l = 1$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_1 = -2$ and $y_1 = 1$.

For $i = 2$, we have $\mathbf{b}_2 = [-5,-3] \times [2,3]$. This box satisfies the conditions $\overline{x}_i = -3 \le x_{(k)} = -2$ and $y_{(l+1)} = 2 \le \underline{y}_{(i)} = 2$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_2 = -5$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [-5,-3] \times [2,3]$. This box satisfies the conditions $\overline{x}_i = -3 \le x_{(k)} = -2$ and $y_{(l+1)} = 2 \le \underline{y}_{(i)} = 2$ Therefore, according to the table from Step 3.1, we take the values $x_i = \underline{x}_i$ and $y_i = \overline{y}_i$ i.e., $x_3 = -5$, $y_3 = 3$.

So, for the zone $\mathbf{z}_{3,2}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (-2,-5,-5); y = (y_1, y_2, y_3) = (1,3,3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{-2-5-5}{3} = -4; E_y = \frac{1+3+3}{3} = \frac{7}{3}.$$

The resulting vector $E^{(0)} = \left(-4, \frac{7}{3}\right)$ does not belong to the box $\mathbf{z}_{3,2} = [-2,-1] \times [1,2]$. So, we dismiss the pair of sequences (x, y).

Zone $\mathbf{z}_{3,3}$. Now, we consider the zone $\mathbf{z}_{3,3} = [-2,-1] \times [2,3]$.

For $i = 1$, we have $\mathbf{b}_1 = [-2,-1] \times [0,1]$. This box satisfies the condition $\underline{x}_i = -2 \le x_{(k)} = -2 \le x_{(k+1)} = -1 \le \overline{x}_i = -1$ and $\overline{y}_i = 1 \le y_{(l)} = 3$ Therefore, according

to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_1 = -1$ and $y_1 = 0$.

For $i = 2$, we have $\mathbf{b}_2 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\overline{x}_i = -3 \leq x_{(k)} = -2$ and $\underline{y}_i = 2 \leq y_{(l)} = 2 \leq y_{(l+1)} = 3 \leq \overline{y}_i = 3$ Therefore, according to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_2 = -5$ and $y_2 = 3$.

For $i = 3$, we have $\mathbf{b}_3 = [-5, -3] \times [2, 3]$. This box satisfies the conditions $\overline{x}_i = -3 \leq x_{(k)} = -2$ and $\underline{y}_i = 2 \leq y_{(l)} = 2 \leq y_{(l+1)} = 3 \leq \overline{y}_i = 3$ Therefore, according to the table from Step 3.1, we take the values $x_i = \overline{x}_i$ and $y_i = \underline{y}_i$ i.e., $x_3 = -5$, $y_3 = 3$.

So, for the zone $\mathbf{z}_{3,3}$, we must consider a single pair of sequences:

$$x = (x_1, x_2, x_3) = (-1, -5, -5); y = (y_1, y_2, y_3) = (0, 3, 3).$$

On Step 3.3, we compute the values of the average for these sequences. We get

$$E_x = \frac{-1 - 5 - 5}{3} = -\frac{11}{3}; E_y = \frac{0 + 3 + 3}{3} = 2.$$

The resulting vector $E^{(0)} = \left(-\frac{11}{3}, 2\right)$ does not belong to the box $\mathbf{z}_{3,3} = [-2, -1] \times [2, 3]$. So, we dismiss the pair of sequences (x, y).

3.5.9 Step 5

On Step 3, we computed a single value of covariance $C_{x,y}$: $-\frac{8}{3}$. According to Step 4, $\underline{C}_{x,y}$ is the smallest of these covariances, so $\underline{C}_{x,y} = -\frac{8}{3}$.

According to Step 5.1, we compute the value of $\overline{C}_{x,y}$ as $-\underline{C}_{-x,y}$, since the value of $\underline{C}_{-x,y}$, is $-\frac{8}{3}$, the value of $\overline{C}_{x,y}$ is $\frac{8}{3}$.

3.5.10 Conclusion

Thus following the above algorithm we get the bounds of the covariance for the intervals, $\mathbf{x}_1 = [1, 2]$, $\mathbf{x}_2 = [3, 5]$, $\mathbf{x}_3 = [3, 5]$, $\mathbf{y}_1 = [0, 1]$, $\mathbf{y}_2 = [2, 3]$, $\mathbf{y}_3 = [2, 3]$, i.e.,

$$\mathbf{C}_{x,y} \in [\underline{C}_{x,y}, \overline{C}_{x,y}] = \mathbf{C}_{x,y} \in \left[\frac{2}{9}, \frac{8}{3}\right].$$

Appendix A

```
//Program to compute the covariance when the
//intervals come from privacy-protected
//statistical databases.

//The input to this program is the number of
//intervals n to be considered followed
//by the intervals x[1],...,x[n],y[1],...,y[n].


//The output to the program is the range of
//covariance.

#include<stdio.h> #include<conio.h>

//Function used for sorting the bounds of x and y intervals
void sort(double Zx[], int n){
    //The input to the function is an array containing the
    //bounds of X- or Y-intervals; for example, for X-intervals,
```

```
    //first interval is [Zx[0], Zx[1]], the second interval is
    //[Zx[2], Zx[3]],..., and the n-th interval is
    //[Zx[2n-2], Zx[2n-1]].
    //In this program, we use selection sort; however, any other
    //sorting algorithm can be used as well.
    int i=0,j=0;//Local variables used for indexing
    double temp;//Local variable used for swapping
    for(i=0;i<2*n;i++){
        for(j=i;j<2*n;j++){
            if(Zx[j]<Zx[i]){
                temp=Zx[i];
                Zx[i]=Zx[j];
                Zx[j]=temp;
            }

         }
    }
}

//Function to eliminate the duplicates in a sorted array.
int EliminateDup(double Zx[],int n){
    double temp[2*n];
    int i=0,ind=0,j=0;
//Storing the array in a temporary array
    for(i=0;i<n;i++)
        temp[i]=Zx[i];
//Eliminating duplicates
```

```
    for(j=0;j<(n-1);j++){
        if(temp[j]==temp[j+1]){
            temp[j]=32760;
        }
    }
    for(i=0;i<n;i++){
        if(temp[i]!=32760){
            Zx[ind]=temp[i];
            ind++;
        }
    }
    return(ind);
}

//Main Program
void main(){
    //Local variables used for indexing the array and
    //to set Flag for specific condition when the
    //corresponding box coincides with the zone.
    int n,i=0,j=0,k=0,l=0,m=0,a=0,index=0,flag=0,n1=0;
    int n2=0,n3=0,n4=0,m1=0;
    int index1=0,b=0,eFlag=0,xn=0,yn=0;
    //Lx[],Ux[],Ly[],Uy[] are used to store the lower
    //and upper bounds of X and Y-intervals
    //Zx[],Zy[] are used to store the sorted interval bounds
    //X[],Y[] are used to store the values of xi and yi
    //where the extrema of the covariance may be attained
```

```
//temp[] is used to store the values of xi and yi when
//the corresponding box coincides with the zone
double Ly[n],Uy[n],Lx[n],Ux[n],Zx[2*n], Zy[2*n];
double X[n],Y[n];
//Variables used to store the averages and
//covariance cxy of the xi and yi values
//Cxy is the smallest of all the covariances cxy
//that we have computed so far.

double Ex=0,Ey=0,cxy,Cxy;

//Initializing the value of covariance Cxy to a large
//value
Cxy=32760;

//Inputting the number of x and y-intervals
printf("Enter The Number of Intervals(n): ");
scanf("%d",&n);


//Inputting the x and y intervals
printf("\n Enter The bounds of X-Interval(Lx[],Ux[]):\n");
for(i=0;i<n;i++){
    scanf("%lf %lf",&Lx[i], &Ux[i]);
    Zx[j]=Lx[i];
    Zx[++j]=Ux[i];
    j=j+1;
```

```
}
j=0;
printf("Enter The bounds of Y-Interval(Ly[],Uy[]):\n");
for(i=0;i<n;i++){
    scanf("%lf %lf",&Ly[i], &Uy[i]);
    Zy[j]=Ly[i];
    Zy[++j]=Uy[i];
    j=j+1;
}

//Sorting the X and Y-intervals
sort(Zx,n);
sort(Zy,n);

//Eliminating duplicates
xn=EliminateDup(Zx,2*n);
yn=EliminateDup(Zy,2*n);

//Dividing the sorted intervals into zones
for(i=0;i<(xn-1);i++){
    for(j=0;j<(yn-1);j++){
        //Checking specific conditions for each zone
        //with respect
        //to each box(formed by Lx[],Ux[],Ly[],Uy[])
        for(k=0;k<n;k++){

            //Checking if the x-upper bound of the box
```

```
//is less than or equal to x-lower bound
//of the zone
if(Ux[k]<=Zx[i]){
    //Checking if the y-upper bound of the
    //box is less than or equal to y-lower
    //bound of the zone
    if(Uy[k]<=Zy[j]){
        X[k]=Ux[k];
        Y[k]=Uy[k];
    }
    //Checking if the y-interval of the zone
    // lies within the y-interval of the box
    if((Ly[k]==Zy[j])&&(Zy[j+1]==Uy[k])){
        X[k]=Lx[k];
        Y[k]=Uy[k];
    }
    //Checking if the y-upper bound of the
    //zone is less than y-lower bound of
    //the box
    if(Zy[j+1]<=Ly[k]){
        X[k]=Lx[k];
        Y[k]=Uy[k];
    }

}

//Checking if the x-interval for the zone
```

```
//is within the x-interval of the box
if((Lx[k]==Zx[i])&&(Zx[i+1]==Ux[k])){
    //Checking if the y-upper bound of the
    //box is less than or equal to y-lower
    //bound of the zone
    if(Uy[k]<=Zy[j]){
        X[k]=Ux[k];
        Y[k]=Ly[k];
    }
    //Checking if the y-interval of the zone
    //lies within the y-interval the box
    if((Ly[k]==Zy[j])&&(Zy[j+1]==Uy[k])){
        //If this condition is met, we set a
        //FLAG and save the index as we have
        // to try all possible combinations
        //of the endpoints.
        if(flag==0)
            index=k;
        flag=1;
        m1++;
        X[k]=0;
        Y[k]=0;

    }
    //Checking if the y-upper bound of the
    //zone is less than y-lower bound of
    //the box
```

```
    if(Zy[j+1]<=Ly[k]){
        X[k]=Lx[k];
        Y[k]=Uy[k];
    }

}

//Checking if the x-upper bound of the zone
//is less than y-lower bound of the box
if(Zx[i+1]<=Lx[k]){
    //Checking if the y-upper bound of the
    //box is less than or equal to y-lower
    //bound of the zone
    if(Uy[k]<=Zy[j]){
        X[k]=Ux[k];
        Y[k]=Ly[k];
    }
    //Checking if the y-interval of the zone
    //lie within the y-interval of the box
    if((Ly[k]==Zy[j])&&(Zy[j+1]==Uy[k])){
        X[k]=Ux[k];
        Y[k]=Ly[k];
    }
    //Checking if the y-upper bound of the
    //zone is less than y-lower bound of
    //the box
    if(Zy[j+1]<=Ly[k]){
```

```
            X[k]=Lx[k];
            Y[k]=Ly[k];
        }

    }

}

//Calculation of the co-variance if the FLAG
//was set, i.e., when the zone coincides with the box
if(flag==1){
    Ex=0;
    Ey=0;
    index1=index;
    for(n1=0;n1<=m1;n1++){
        n2=m1-n1;
        index=index1;

        for(a=0;a<n1;a++){
            X[index]=Lx[index1];
            Y[index]=Ly[index1];
            index++;
        }
        for(a=0;a<n2;a++){
            X[index]=Ux[index1];
            Y[index]=Uy[index1];
            index++;
```

```
}
for(l=0;l<n;l++){
    Ex=(double)(X[l]+Ex);
    Ey=(double)(Y[l]+Ey);
}

Ex=(double)(Ex/n);
Ey=(double)(Ey/n);

//Initializing values for the
//calculation of covariance
cx=0;
cy=0;
cxy=0;

//Checking to see if the average lies
//within the zone and
//if so calculating covariance

if((Ex>=Zx[i])&&(Ex<=Zx[i+1])
   &&(Ey>=Zy[j])&&(Ey<=Zy[j+1])){
    for(m=0;m<k;m++){
        cxy=(double)(cxy+ ((X[m]-Ex)
            *(Y[m]-Ey)));
    }
    cxy=(double)(cxy/n);
    //Checking for the value of minimum
```

```
            //covariance
            if(cxy<Cxy){
                Cxy=(double)cxy;
            }
        }
        Ex=0;
        Ey=0;
    }

    flag=0;
    m1=0;
}


//Calculating covariance if the FLAG is not set
//i.e., zone doesn't lie within the box
else{
    //Initializing the x-average and y-average to
    //zero
    Ex=0;
    Ey=0;
    //Calculating the average for X and Y values
    for(l=0;l<n;l++){
        Ex=(double)(X[l]+Ex);
        Ey=(double)(Y[l]+Ey);
    }
    Ex=(double)(Ex/n);
```

```
            Ey=(double)(Ey/n);
            //Initializing values for the calculation of
            //covariance
            cx=0;
            cy=0;
            cxy=0;
            //Checking to see if the average lies within
            //the zone and if
            //so calculating covariance
            if((Ex>=Zx[i])&&(Ex<=Zx[i+1])&&(Ey>=Zy[j])
                &&(Ey<=Zy[j+1])){
                for(m=0;m<k;m++){
                  cxy=(double)(cxy+ ((X[m]-Ex)*(Y[m]-Ey)));
                }
                cxy=(double)(cxy/n);
                //Checking for the value of minimum
                //covariance
                if(cxy<Cxy)
                    Cxy=(double)cxy;
            }
        }
    }

}

printf("The Value of Covariance is :%lf\n", Cxy);
```

```
}
```

References

[1] Beck, J., Kreinovich, V., and B. Wu, Interval-Valued and Fuzzy-Valued Random Variables: From Computing Sample Variances to Computing Sample Covariances, Lopez, M., Gil, M. A., Grzegorzewski, P., Hrynewicz, O., and J. Lawry (eds.), *Soft Methodology and Random Information Systems*, Springer-Verlag, 2004, pp. 85–92

[2] Berleant, D., Automatically verified arithmetic with both intervals and probability density functions, *Interval Computations*, 1993, (2):48–70.

[3] Berleant, D., Automatically verified arithmetic on probability distributions and intervals, In: Kearfott, R. B., and V. Kreinovich, editors, *Applications of Interval Computations*, Kluwer, Dordrecht, 1996.

[4] Berleant, D., and C. Goodman-Strauss, Bounding the results of arithmetic operations on random variables of unknown dependency using intervals, *Reliable Computing*, 1998, 4(2):147–165.

[5] Berleant, D., Xie, L., and J. Zhang, Statool: A Tool for Distribution Envelope Determination (DEnv), an Interval-Based Algorithm for Arithmetic on Random Variables, *Reliable Computing*, 2003, 9(2):91–108.

[6] Denning, D. E. R. D., *Cryptography and data security*, Addison-Wesley, Reading, MA, 1982.

[7] Ferson, S. *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, CRC Press, Boca Raton, Florida, 2002.

[8] Ferson, S., Ginzburg, L., Kreinovich, V., Longpré, L., and M. Aviles, Computing Variance for Interval Data is NP-Hard, *ACM SIGACT News*, 2002, 33(2):108–118.

[9] Ferson, S., Myers, D., and D. Berleant, *Distribution-free risk analysis: I. Range, mean, and variance*, Applied Biomathematics, Technical Report, 2001.

[10] Jaulin, L., Kieffer, M., Didrit, O., and Walter, E. *Applied Interval Analysis*, Springer-Verlag, Berlin, 2001.

[11] Kreinovich, V., Lakeyev, A., Rohn, J., and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1997.

[12] Kreinovich, V., and L. Longpré, Computational complexity and feasibility of data processing and interval computations, with extension to cases when we have partial information about probabilities, In: Brattka, V., Schroeder, M., Weihrauch, K., and N. Zhong, editors, *Proc. Conf. on Computability and Complexity in Analysis CCA'2003*, Cincinnati, Ohio, USA, August 28–30, 2003, pp. 19–54.

[13] Kuznetsov, V. P., *Interval Statistical Models*, Radio i Svyaz, Moscow, 1991 (in Russian).

[14] Lodwick, W. A., and Jamison, K. D., Estimating and Validating the Cumulative Distribution of a Function of Random Variables: Toward the Development of Distribution Arithmetic, *Reliable Computing*, 2003, 9(2):127–141.

[15] Modave, F., Kreinovich, V., Xiang, G., Beck, J., Tupelly, K., Kandathi, R., Longpré, L., Villaverde, K., Debroux, P., and J. Boehm, Using 1-D Radar Observations to Detect a Space Explosion Core Among the Explosion Fragments: Sequential and Distributed Algorithms, *Proceedings of the 11th IEEE Digital Signal Processing Workshop*, Taos, New Mexico, August 1–4, 2004, pp. 273–277.

[16] Moore, R. E., and W. A. Lodwick, Interval Analysis and Fuzzy Set Theory, *Fuzzy Sets and Systems*, 2003, 135(1):5–9.

[17] Nivlet, P., Fournier, F., and J. Royer, A new methodology to account for uncertainties in 4-D seismic interpretation, *Proc. 71st Annual Int'l Meeting of Soc. of Exploratory Geophysics SEG'2001*, San Antonio, TX, September 9–14, 2001, 1644–1647.

[18] Nivlet, P., Fournier, F., and J. Royer, Propagating interval uncertainties in supervised pattern recognition for reservoir characterization, *Proc. 2001 Society of Petroleum Engineers Annual Conf. SPE'2001*, New Orleans, LA, September 30–October 3, 2001, paper SPE-71327.

[19] Osegueda, R., Kreinovich, V., Potluri, L., and R. Aló, Non-Destructive Testing of Aerospace Structures: Granularity and Data Mining Approach. In: *Proc. FUZZ-IEEE'2002*, Honolulu, HI, May 12–17, 2002, Vol. 1, pp. 685–689

[20] Rabinovich, S., *Measurement Errors: Theory and Practice*. American Institute of Physics, New York, 1993.

[21] Regan, H., Ferson, S., and Berleant. D, Equivalence of five methods for bounding uncertainty, *International Journal of Approximate Reasoning*, 2004, 36:1–30.

[22] Rowe, N. C., Absolute bounds on the mean and standard deviation of transformed data for constant-sign-derivative transformations, *SIAM Journal of Scientific Statistical Computing*, 1988, 9:1098–1113.

[23] Starks, S. A., Kreinovich, V., Longpré, L., Ceberio, M., Xiang, G., Araiza, R., Beck, J., Kandathi, K., Nayak, A., and R. Torres, Towards Combining Probabilistic and Interval Uncertainty in Engineering Calculations, *Proceedings of the Workshop on Reliable Engineering Computing*, Savannah, Georgia, September 15–17, 2004, pp. 193–213.

[24] Walley, P., *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, N.Y., 1991.

[25] Williamson, R., and T. Downs, Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds, *International Journal of Approximate Reasoning*, 1990, 4:89–158.

[26] Xiang, G., Starks, S. A., Kreinovich, V., and L. Longpré, New Algorithms for Statistical Analysis of Interval Data, *Proceedings of the Workshop on State-of-the-Art in Scientific Computing PARA'04*, Lyngby, Denmark, June 20–23, 2004, Vol. 1, pp. 123–129.

Curriculum Vitae

Raj K. Kandathi, the eldest son of Venu Gopal and Sarada, was born on April 10, 1979 in Nellore, Andhra Pradesh, India. He graduated from the Jawaharlal Nehru Technical University, Hyderbad, with a Bachelor degree of Engineering in Computer Science and Engineering in the Spring of 2002.

In the Fall of 2002, Raj came to the University of Texas at El Paso for his Master's degree in Computer Science. While pursuing graduate studies, he worked as a Tutor in the El Paso Community College, Teaching Assistant for the Computer Science Department, and as a Computer Programmer for Datamark Inc.

During his graduate studies, he co–authored the following two publications:

- Scott A. Starks, Vladik Kreinovich, Luc Longpré, Martine Ceberio, Gang Xiang, Roberto Araiza, Jan Beck, Raj Kandathi, Aziz Nayak, Roberto Torres, "Towards Combining Probabilistic and Interval Uncertainty in Engineering Calculations", Proceedings of the Workshop on Reliable Engineering Computing, Savannah, Georgia, September 15–17, 2004, pp. 193–213.

- F. Modave, V. Kreinovich, G. Xiang, J. Beck, K. Tupelly, R. Kandathi, L. Longpré, K. Villaverde, P. Debroux, J. Boehm, "Using 1-D Radar Observations to Detect a Space Explosion Core Among the Explosion Fragments: Sequential and Distributed Algorithms", Proceedings of the 11th IEEE Digital Signal Processing Workshop, Taos, New Mexico, August 1–4, 2004, pp. 273–277.

Permanent address: Plot # 37, M.I.G.
Law Sons Bay Colony, Vizag – 530017
Andhra Pradesh, India.

This thesis was typed by the author.