5-1-2012

Estimating Correlation under Interval and Fuzzy Uncertainty: Case of Hierarchical Estimation

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

Comments:
Technical Report: UTEP-CS-12-17
Published in *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2012*, Berkeley, California, August 6-8, 2012.

Estimating Correlation under Interval and Fuzzy Uncertainty: Case of Hierarchical Estimation

Ali Jalal-Kamali
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968
Email: ajalalkamali@miners.utep.edu

***Abstract*—In many situations, we are interested in finding the correlation ρ between different quantities x and y based on the values x_i and y_i of these quantities measured in different situations i. The correlation is easy to compute when we know the exact sample values x_i and y_i. In practice, the sample values come from measurements or from expert estimates; in both cases, the values are not exact. Sometimes, we know the probabilities of different values of measurement errors, but in many cases, we only know the upper bounds Δ_{xi} and Δ_{yi} on the corresponding measurement errors. In such situations, after we get the measurement results $\widetilde{x}_i$ and $\widetilde{y}_i$, the only information that we have about the actual (unknown) values x_i and y_i is that they belong to the corresponding intervals $[\widetilde{x}_i - \Delta_{xi}, \widetilde{x}_i + \Delta_{xi}]$ and $[\widetilde{y}_i - \Delta_{yi}, \widetilde{y}_i + \Delta_{yi}]$. For expert estimates, we get different intervals corresponding to different degrees of certainty – i.e., fuzzy sets. Different values of x_i and y_i lead, in general, to different values of the correlation ρ. It is therefore desirable to find the range $[\underline{\rho}, \overline{\rho}]$ of possible values of the correlation when x_i and y_i take values from the corresponding intervals. In general, the problem of computing this range is NP-hard. In this paper, we provide a feasible (= polynomial-time) algorithm for computing at least one of the endpoints of this interval: for computing $\overline{\rho}$ when $\overline{\rho} > 0$ and for computing $\underline{\rho}$ when $\underline{\rho} < 0$.**

I. Introduction

Need for correlation. In many practical situations, it is desirable to know which quantities are independent and which are correlated – positively or negatively.

To estimate the correlation between the quantities x and y, we repeatedly measure the values x_i and y_i of both quantities in different situations i. The correlation ρ is then estimated as the ratio $\rho = \dfrac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}$, of the covariance C to the product of standard deviations $\sqrt{V_x}$ and $\sqrt{V_y}$. Covariance and standard deviations, in their turn, are defined as follows:

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

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

and the means E_x and E_y are estimated as follows:

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

Need to take into account interval uncertainty. The values x_i and y_i used to estimate correlation come from measurements, and measurements are never absolutely accurate: the measurement results $\widetilde{x}_i$ and $\widetilde{y}_i$ are, in general, different from the actual (unknown) values x_i and y_i of the corresponding quantities. As a result, the value $\widetilde{\rho}$ estimated based on these measurement results is, in general, different from the ideal value ρ which we would get if we could use the actual values x_i and y_i. It is therefore desirable to determine how accurate is the resulting estimate.

Sometimes, we know the probabilities of different values of measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ and $\Delta y_i \stackrel{\text{def}}{=} \widetilde{y}_i - y_i$. However, in many cases, we do not know these probabilities, we only know the upper bounds Δ_{xi} and Δ_{yi} on the corresponding measurement errors: $|\Delta x_i| \le \Delta_{xi}$ and $|\Delta y_i| \le \Delta_{yi}$; see, e.g., [5]. In this case, the only information that we have about the actual values x_i and y_i is that they belong to the corresponding intervals $[\underline{x}_i, \overline{x}_i] = [\widetilde{x}_i - \Delta_{xi}, \widetilde{x}_i + \Delta_{xi}]$ and $[\underline{y}_i, \overline{y}_i] = [\widetilde{y}_i - \Delta_{yi}, \widetilde{y}_i + \Delta_{yi}]$. Different values $x_i \in [\underline{x}_i, \overline{x}_i]$ and $y_i \in [\underline{y}_i, \overline{y}_i]$ lead, in general, to different values of the covariance. It is therefore desirable to find the range of all possible values of the covariance ρ:

$$[\underline{\rho}, \overline{\rho}] =$$
$$\{\rho(x_1, \ldots, x_n, y_1, \ldots, y_n) : x_i \in [\underline{x}_i, \overline{x}_i], y_i \in [\underline{y}_i, \overline{y}_i]\}.$$

Case of expert uncertainty, and how the corresponding computations can be reduced to the interval case. An expert usually describes his/her uncertainty by using words from the natural language, like "most probably, the value of the quantity is between 6 and 7, but it is somewhat possible to have values between 5 and 8". To formalize this knowledge, it is natural to use fuzzy set theory, in which, for every value x_i, we have a fuzzy set $\mu_i(x_i)$ which describes the expert's knowledge about x_i. An alternative user-friendly way to represent a fuzzy set is by using its α-cuts $\mathbf{x}_i(\alpha) = \{x_i \,|\, \mu_i(x_i) \ge \alpha\}$. It is known (see, e.g., [4]) that for any function $y = f(x_1, \ldots, x_n)$, the α-cut of y is equal to

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

Thus, from the computational viewpoint, the problem of estimating ρ under fuzzy uncertainty can be reduced to several

similar problems for interval uncertainty – interval problems corresponding to different values α. In view of this reduction, in the following text, we will concentrate on estimating the correlation under interval uncertainty.

What is known. The problem of estimating correlation under interval uncertainty is, in general, NP-hard [3]. This means, crudely speaking, that unless P=NP (which most computable scientists believe to be impossible), no feasible (i.e., no polynomial-time) algorithm is possible that would always compute the range of the corresponding characteristic under interval uncertainty.

In [1], we showed that while we cannot have an efficient algorithm for computing both bounds $\underline{\rho}$ and $\overline{\rho}$, we can effectively compute (at least) one of the bounds. We consider the case that we have weights w_i for corresponding values x_i and y_i and we show that our claim still holds. Specifically, that we can compute $\overline{\rho}$ when $\overline{\rho} > 0$ and we can compute $\underline{\rho}$ when $\underline{\rho} < 0$. This means that, in the case of a non-degenerate interval $[\underline{\rho}, \overline{\rho}]$ (i.e., $\underline{\rho} < \overline{\rho}$):

- when $\overline{\rho} \le 0$, we compute the lower endpoint $\underline{\rho}$;
- when $0 \le \underline{\rho}$, we compute the upper endpoint $\overline{\rho}$;
- in all remaining cases, when $\underline{\rho} < 0 < \overline{\rho}$, we compute both lower endpoint $\underline{\rho}$ and $\overline{\rho}$.

Need to take into account that the estimation is usually hierarchical. In some practical situations, e.g., when processing the census results, we do not process all the census data, what we usually do is we first combine the data by town, then combine town data into state-wide date, etc.; see, e.g., [2], [6].

In general, on each stage, the data points are divided into groups $I_1, \ldots, I_m$, and instead of directly processing all the data points, we process the results of previous processing within each of these m groups. For example, in the previous processing, we have compute the averages $E_{xj} = \dfrac{1}{n_j} \cdot \sum\limits_{i \in I_j} x_i$ over each group. Now, the overall average E_x can be described as

$$E_x = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot \sum_{j=1}^{m} \sum_{i \in I_j} x_i = \sum_{j=1}^{m} p_j \cdot E_{xj},$$

i.e.,

$$E_x = \sum_{j=1}^{m} p_j \cdot E_{xj}, \tag{1}$$

where we denoted $p_j \stackrel{\text{def}}{=} \dfrac{n_j}{n}$. Similarly [6],

$$E_y = \sum_{j=1}^{m} p_j \cdot E_{yj}, \tag{2}$$

$$V_x = \sum_{j=1}^{m} p_j \cdot (E_{xj} - E_x)^2 + \sum_{j=1}^{m} p_j \cdot V_{xj}, \tag{3}$$

$$V_y = \sum_{j=1}^{m} p_j \cdot (E_{yj} - E_y)^2 + \sum_{j=1}^{m} p_j \cdot V_{yj}, \tag{4}$$

where V_{xj} and V_{yj} are sample variances within the corresponding group. For covariance, we have

$$C = \sum_{j=1}^{m} p_j \cdot (E_{xj} - E_x) \cdot (E_{yj} - E_y) + \sum_{j=1}^{m} p_j \cdot C_j, \tag{5}$$

where C_j is the covariance over each group. Finally, we compute correlation ρ as

$$\rho = \frac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}. \tag{6}$$

Hierarchical estimation under interval uncertainty. In the ideal case, for each group j, we know the values p_j, E_{xj}, E_{yj}, V_{xj}, V_{yj}, and C_j. Based on these values, we compute E, V_x, V_y, C, and finally, the correlation ρ. In practice, we often only know the values x_i and y_i with interval uncertainty. As a result, for each group j, instead of the exact value of the each of the above characteristics, we only know the interval of its possible values, i.e., we know the intervals $\mathbf{E}_{xj}$, $\mathbf{E}_{xj}$, $\mathbf{E}_{yj}$, $\mathbf{V}_{xj}$, $\mathbf{V}_{yj}$, and $\mathbf{C}_j$. Different values from these intervals lead, in general, to different correlation values ρ. It is therefore desirable to find the range $[\underline{\rho}, \overline{\rho}]$ of possible values of the correlation obtained by formulas (1)-(6).

What we do in this paper. In this paper, we show that for such a hierarchical estimation, it is still possible to feasibly compute at least one of the endpoints of the interval of possible values of the correlation ρ.

II. Main Result and the Corresponding Algorithm

Main Result. *There exists a polynomial-time algorithm that, given m numbers p_j and m groups of intervals $\mathbf{E}_{xj}$, $\mathbf{E}_{xj}$, $\mathbf{E}_{yj}$, $\mathbf{V}_{xj}$, $\mathbf{V}_{yj}$, and $\mathbf{C}_j$, computes (at least) one of the endpoint of the interval $[\underline{\rho}, \overline{\rho}]$ of possible values of the correlation ρ:*

- *it computes $\overline{\rho}$ if $\overline{\rho} > 0$, and*
- *it computes $\underline{\rho}$ if $\underline{\rho} < 0$.*

Reducing minimum to maximum. One can easily see that when we change the signs of E_{yj} and C_j, the correlation changes sign as well. It is known that

$$\min f(x) = -\max(-f(x)).$$

So, if if we know how to compute the largest value $\overline{\rho}$ when this value is positive, we can then compute the smallest value $\underline{\rho}$ when this value is negative, by taking intervals $\mathbf{E}'_{yj} = -\mathbf{E}_{yj}$ and $\mathbf{C}'_j = -\mathbf{C}_j$, computing the corresponding bound $\overline{\rho'}$, and then taking $\underline{\rho} = -\overline{\rho'}$.

Because of this reduction, in the following text, we will concentrate on computing the largest value $\overline{\rho}$.

Preliminary observation. In the ratio ρ, the dependence on C_j is only in the numerator, and the dependence on V_{xj} and V_{yj} is only in the denominator. Thus, the ratio ρ is the largest when each term C_j attains its largest possible value $\overline{C}_j$, and when each term V_{xj} and V_{yj} attains its smallest possible value $\underline{V}_{xj}$ and $\underline{V}_{yj}$. So, in the following text, we will take $C_j = \overline{C}_j$,

$V_{xj} = \underline{V}_{xj}$, and $V_{yj} = \underline{V}_{yj}$, and consider only the dependence on X_{xj} and E_{yj}.

Algorithm. For each j from 1 to m, the corresponding box $[\underline{E}_{xj}, \overline{E}_{xj}] \times [\underline{y}_i, \overline{E}_{yj}]$ has four vertices: $(\underline{E}_{xj}, \underline{E}_{yj})$, $(\underline{E}_{xj}, \overline{E}_{yj})$, $(\overline{E}_{xj}, \underline{E}_{yj})$, and $(\overline{E}_{xj}, \overline{E}_{yj})$. So, totally, we have $4n$ vertices.

Let us consider all 4-tuples consisting of two vertices and two signs. For each pair of vertices, there are nine possible combinations of two $+$, $-$, or 0 signs: $(-,-)$, $(-,0)$, $(-,+)$, $(0,-)$, $(0,0)$, $(0,+)$, $(+,-)$, $(+,0)$, and $(+,+)$.

For each 4-tuple, if the first sign is not 0, we move the first vertex slightly along the x axis in the direction determined by the first sign, i.e.:

- slightly increase x if the sign is $+$ and
- slightly decrease x if the sign is $-$.

Here, "slightly" means that the change is much smaller than the smallest difference between distinct values E_{xj} and E_{yj}.

Then, if the second sign is not 0, we move the second vertex slightly along the x axis in the direction determined by the second sign. Thus, we get two points on the (x, y) plane. We can then form a straight line going through these two points.

Now, we select two 4-tuples, and form two lines. We will call the first line *representative x-line*, and the second line *representative y-line*.

If we selected the same line as the representative x-line and the representative y-line, then we check whether this line intersects each of n boxes. If it does, then $\overline{\rho} = 1$. If this line does not have a common point with one of the boxes, we dismiss this selection, and continue with other selections.

Let us explain the algorithm in the cases when the representative x-line and the representative y-line are different. The representative x-line divides the plane into two semi-planes:

- the points *above* this line, i.e., the points (x, y) for which the y coordinate is larger than the y-value of the point on the x-line with the same x coordinate, and
- the points *below* this line, i.e., the points (x, y) for which the y coordinate is smaller than the y-value of the point on the x-line with the same x coordinate.

The representative y-line similarly divides the plane into two semi-planes:

- the points to the *right* of this line, i.e., the points (x, y) for which the x coordinate is larger than the x-value of the point on the x-line with the same y coordinate, and
- the points to the *left* of this line, i.e., the points (x, y) for which the x coordinate is smaller than the x-value of the point on the y-line with the same y coordinate.

Based on where each of the vertices is with respect to these two lines, we can tell the relation of each box $[\underline{E}_{xj}, \overline{E}_{xj}] \times [\underline{E}_{yj}, \overline{E}_{yj}]$ with respect to each line.

The lines that we computed are "representatives" of the actual lines that we will be using, in the sense that the actual lines will have the exact same relation to each of the n boxes. Let us describe the corresponding *actual* lines as follows:

- the actual x-line has the form $y = E_y + k_x \cdot (x - E_x)$, and
- the actual y-line has the form $x = E_x + k_y \cdot (y - E_y)$,

where E_x, E_y, k_x, and k_y are to-be-determined real numbers.

For each box $[\underline{E}_{xj}, \overline{E}_{xj}] \times [\underline{E}_{yj}, \overline{E}_{yj}]$, based on its location in comparison to the representative lines, we select the values E_{xj} and E_{yj} as follows:

- If the whole box is above the representative x-line, we take $E_{xj} = \overline{E}_{xj}$. On the resulting segment $\{\overline{E}_{xj}\} \times [\underline{E}_{yj}, \overline{E}_{yj}]$, we select the point which is the closest to the actual y-line:
  - if the whole segment is to the right of the representative y-line, we select $E_{yj} = \underline{E}_{yj}$;
  - if the whole segment is to left of the representative y-line, we select $E_{yj} = \overline{E}_{yj}$;
  - if the segment intersects with the representative y-line, we select the value E_{yj} corresponding to the intersection point between the segment and the actual y-line.
- If the whole box is below the representative x-line, we take $E_{xj} = \underline{E}_{xj}$. On the resulting segment $\{\underline{E}_{xj}\} \times [\underline{E}_{yj}, \overline{E}_{yj}]$, we select the point which is the closest to the actual y-line:
  - if the whole segment is to the right of the representative y-line, we select $E_{yj} = \underline{E}_{yj}$;
  - if the whole segment is to left of the representative y-line, we select $E_{yj} = \overline{E}_{yj}$;
  - if the segment intersects with the representative y-line, we select the value E_{yj} corresponding to the intersection point between the segment and the actual y-line.
- If the whole box is to the right of the representative y-line, we take $E_{yj} = \underline{E}_{yj}$. On the resulting segment $[\underline{E}_{xj}, \overline{E}_{xj}] \times \{\underline{E}_{yj}\}$, we select the point which is the closest to the actual x-line:
  - if the whole segment is above the representative x-line, we select $E_{xj} = \underline{E}_{xj}$;
  - if the whole segment is below the representative x-line, we select $E_{xj} = \overline{E}_{xj}$;
  - if the segment intersects with the representative x-line, we select the value E_{xj} corresponding to the intersection point between this segment and the actual x-line.
- If the whole box is to the left of the representative y-line, we take $E_{yj} = \overline{E}_{yj}$. On the resulting segment $[\underline{E}_{xj}, \overline{E}_{xj}] \times \{\overline{E}_{yj}\}$, we select the point which is the closest to the actual x-line:
  - if the whole segment is above the representative x-line, we select $E_{xj} = \underline{E}_{xj}$;
  - if the whole segment is below the representative x-line, we select $E_{xj} = \overline{E}_{xj}$;
  - if the segment intersects with the representative x-line, we select the value E_{xj} corresponding to the intersection point between the segment and the actual x-line.
- The only remaining case is when the box contains the intersection point (E_x, E_y) of the actual x- and y-lines.

Thus, for each j and for each of the values E_{xj} and E_{yj}, we get an explicit expression in terms of the four parameters E_x, E_y, k_x and k_y (the parameters that describe the actual x- and y-lines).

By substituting these expressions for E_{xj} and E_{yj} into the following formulas, we get a system of four equations with four unknowns E_x, E_y, k_x and k_y:

$$E_x = \sum_{j=1}^{m} p_j \cdot E_{xj}; \tag{7}$$

$$E_y = \sum_{j=1}^{m} p_j \cdot E_{yj}; \tag{8}$$

$$\sum_{i=1}^{n} p_j \cdot E_{xj} \cdot E_{yj} - E_x \cdot E_y + \sum_{j=1}^{m} p_j \cdot \overline{C}_j = k_x \cdot \left(\sum_{j=1}^{m} p_j \cdot (E_{xj} - E_x)^2 + \sum_{j=1}^{m} p_j \cdot \underline{V}_{xj} \right); \tag{9}$$

$$\sum_{j=1}^{n} p_j \cdot E_{xj} \cdot E_{yj} - E_x \cdot E_y + \sum_{j=1}^{m} p_j \cdot \overline{C}_j = k_y \cdot \left(\sum_{j=1}^{m} p_j \cdot (E_{yj} - E_y)^2 + \sum_{j=1}^{m} p_j \cdot \underline{V}_{yj} \right). \tag{10}$$

Once we solve this system, we get one or several possible solutions. For each of these solutions, we can form the corresponding actual x- and y-lines.

Then, we check whether each of $4n$ vertices is in the same relation to the resulting two lines and to the representative x- and y-lines, i.e., e.g., that each vertex is above, below, or on the actual x-line if and only if it is, correspondingly, above, below, or on the corresponding representative x-line, and that the same property holds for the y-lines. If at least one vertex is in a different relation, we dismiss this solution. Otherwise, we compute the value of the correlation ρ based on the corresponding values E_{xj} and E_{yj}.

The largest of all the values ρ corresponding to all possible pairs of tuples is then returned as the desired value $\overline{\rho}$.

Comment. For each pair of lines, for each j, according to our algorithm, as the appropriate value of E_{xj}, we make one of the following four selections:

- sometimes, we select a known value $\underline{E}_{xj}$;
- sometimes, we select a know value $\overline{E}_{xj}$;
- sometimes, we select the value $E_{xj} = E_x$ (which is not a priori known, it is one of the four variables that we need to determine), and
- sometimes, we select a value E_{xj} that lies on the x-line $y = E_y + k_x \cdot (E_{xj} - E_x)$, i.e., a value

$$E_{xj} = E_x + K_x \cdot (E_{yj} - E_y),$$

where $K_x \stackrel{\text{def}}{=} \dfrac{1}{k_x} = \dfrac{V_x}{C}$.

In general, each expression E_{xj} is a linear combination of a constant and the unknowns E_x, K_x, and $K_x \cdot E_y$. According to the algorithm, for each i, it takes a finite number of computational steps to check the corresponding conditions and, based on the results of this checking, to find the appropriate value E_{xj}. Similarly, each expression E_{yj} is a linear combination of a constant and the unknowns E_y, K_y, and $K_y \cdot E_x$.

Substituting these expressions for E_{xj} and E_{yj} into the four equations for the unknowns E_x, E_y, K_x, and K_y, we conclude that:

- the equation (7) for E_x is transformed into equating a linear combination of E_x, K_x, and $K_x \cdot E_y$ to zero;
- the equation (8) for E_y is transformed into equating a linear combination of E_y, K_y, and $K_y \cdot E_x$, to zero;
- the equations (9) and (10) (corresponding to $k_x \cdot V_x = k_y \cdot V_y = C$) are transformed into equating a linear combination of terms of order ≤ 4 in terms of the unknowns.

As a result, to find the four unknown E_x, E_y, K_x, and K_y, we get a system of four polynomial equations of order ≤ 4. The amount of computation time which is needed to solve this system does not depend on the size m of the input, so in terms of dependence on this size, we need $O(1)$ time.

III. Proof of the Main Result

Proof that the above algorithm is polynomial time. Before we prove that the algorithm is correct, let us first prove that it is indeed a polynomial time algorithm.

We have $4m$ possible vertices, so we have $O(m^2)$ possible pairs of vertices – and thus, $O(m^2)$ possible 4-tuples. Thus, we have $O(m^2)$ possible representative x-lines, and we also have $O(m^2)$ representative y-lines. In our algorithms, we consider pairs consisting of a representative x-line and a representative y-line. Since we have $O(m^2)$ x-lines and we have $O(m^2)$ y-lines, we therefore have $O(m^2) \cdot O(m^2) = O(m^4)$ possible pairs consisting of a representative x-line and a representative y-line.

For each pair of lines, we perform the following computations:

- First, need a constant number of steps to find the expression for each of m values E_{xj} and each of n values E_{yj} in terms of the parameters E_x, E_y, K_x, and K_y. So, we need $O(m)$ steps to find these expressions for all j.
- Then, we need linear time $O(m)$ to form the corresponding systems of four equations with four unknowns and constant time $O(1)$ to solve this system.
- Once this system is solved, and we know the corresponding values E_x, E_y, k_x, and k_y, we need:
  - a linear time $O(m)$ to check whether each of $4m = O(m)$ vertices is in the right position with respect to the corresponding lines, and,
  - if needed, linear time $O(m)$ to compute the corresponding value of the correlation ρ – by using the above explicit formulas (1)-(6) describing how the correlation ρ depends on E_{xj} and E_{yj}.

Totally, for each pair of lines, we need

$$O(m) + O(m) + O(1) + O(m) + O(m) = O(m)$$

computational steps.

We need $O(m)$ steps for each of $O(m^4)$ pairs of lines. Thus, the total computation time of this algorithm is $O(m^4) \cdot O(m) = O(m^5)$ – which is indeed polynomial in the size m of the problem.

Case when the representative x-line coincides with the representative y-line. If this common line intersects with all m boxes $[\underline{E}_{xj}, \overline{E}_{xj}] \times [\underline{E}_{yj}, \overline{E}_{yj}]$, then, for each box, we can select values E_{xj} and E_{yj} for which the corresponding point (E_{xj}, E_{yj}) belongs to this line. Then, all selected values (E_{xj}, E_{yj}) follow the same linear dependence $E_{yj} = E_y + k_x \cdot (E_{xj} - E_x)$ (as described by the common lines). Therefore, for this selection, the correlation is 1. Since $\rho \le 1$, this means that in this case, $\overline{\rho} = 1$.

Remaining cases. Let us now prove that our algorithm is correct for all other cases, when the x- and the y-lines are different.

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

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

Computing the corresponding derivatives. Based on the expression we had for E_x, we conclude that $\dfrac{\partial E_x}{\partial E_{xj}} = p_j$ and similarly $\dfrac{\partial E_y}{\partial E_{yj}} = p_j$. Since the variance V_x can be described in an equivalent form

$$V_x = \sum_{j=1}^{m} p_j X_{xj}^2 - E_x^2 + \sum_{j=1}^{m} p_j \cdot \underline{V}_{xj},$$

we get $\dfrac{\partial V_x}{\partial E_{xj}} = 2p_j \cdot (E_{xj} - E_x)$. Similarly, we get $\dfrac{\partial V_y}{\partial E_{yj}} = 2 \cdot p_j \cdot (E_{yj} - E_y)$. The covariance can be equivalently rewritten as

$$C = \sum_{j=1}^{m} p_j \cdot E_{xj} \cdot E_{yj} - E_x \cdot E_y + \sum_{j=1}^{m} p_j \cdot \overline{C}_j,$$

hence

$$\frac{\partial C}{\partial E_{xj}} = p_j \cdot (E_{yj} - E_y) \text{ and } \frac{\partial C}{\partial E_{yj}} = p_j \cdot (E_{xj} - E_x).$$

So, for $\rho = \dfrac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}$, we get

$$\frac{\partial \rho}{\partial E_{xj}} = \frac{p_j}{\sqrt{V_y} \cdot V_x} \cdot \left[(E_{yj} - E_y) \cdot \sqrt{V_x} - C \cdot \frac{E_{xj} - E_x}{\sqrt{V_x}} \right].$$

Since the standard deviations are always non-negative, the sign of this derivative coincides with the sign of the value

$$(E_{yj} - E_y) \cdot \sqrt{V_x} - C \cdot \frac{E_{xj} - E_x}{\sqrt{V_x}}.$$

Dividing this expression by a positive value $\sqrt{V_x}$, we conclude that the sign of the derivative $\dfrac{\partial \rho}{\partial E_{xj}}$ coincides with the sign of the expression $(E_{yj} - E_y) - k_x \cdot (E_{xj} - E_x)$, where we denoted $k_x \stackrel{\text{def}}{=} \dfrac{C}{V_x}$.

Similarly, the sign of the derivative $\dfrac{\partial \rho}{\partial E_{yj}}$ coincides with the sign of the expression $(E_{xj} - E_x) - k_y \cdot (E_{yj} - E_y)$, where we denoted $k_y \stackrel{\text{def}}{=} \dfrac{C}{V_y}$.

Let us apply the known facts from calculus to this situation. Let E_{xj} and E_{yj} be the values from the corresponding boxes for which the correlation ρ attains its largest possible value $\overline{\rho} > 0$. Then, according to the above facts from calculus, we have one of the three possible situations:

- $E_{xj} \in (\underline{E}_{xj}, \overline{E}_{xj})$ and $\dfrac{\partial \rho}{\partial E_{xj}} = 0$, i.e.,

$$E_{yj} = E_y + k_x \cdot (E_{xj} - E_x);$$

- $E_{xj} = \underline{E}_{xj}$ and $\dfrac{\partial \rho}{\partial E_{xj}} \le 0$, i.e.,

$$E_{yj} \le E_y + k_x \cdot (E_{xj} - E_x);$$

- $E_{xj} = \overline{E}_{xj}$ and $\dfrac{\partial \rho}{\partial E_{xj}} \ge 0$, i.e.,

$$E_{yj} \ge E_y + k_x \cdot (E_{xj} - E_x).$$

Here, k_x has the same sign as the correlation, so $k_x > 0$. Let us now consider possible locations of the box $[\underline{E}_{xj}, \overline{E}_{xj}] \times [\underline{E}_{yj}, \overline{E}_{yj}]$ with respect to the x-line

$$E_{yj} = E_y + k_x \cdot (E_{xj} - E_x).$$

1°. The first case is when the whole box $[\underline{E}_{xj}, \overline{E}_{xj}] \times [\underline{E}_{yj}, \overline{E}_{yj}]$ is above the x-line $E_{yj} = E_y + k_x \cdot (E_{xj} - E_x)$, i.e., when $E_{yj} > E_y + k_x \cdot (E_{xj} - E_x)$ for all $E_{yj} \in [\underline{E}_{yj}, \overline{E}_{yj}]$ and $E_{xj} \in [\underline{E}_{xj}, \overline{E}_{xj}]$. In this case, we cannot have $E_{xj} \in (\underline{E}_{xj}, \overline{E}_{xj})$ and $E_{xj} = \underline{E}_{xj}$, so we must have $E_{xj} = \overline{E}_{xj}$.

On the segment $E_{xj} = \overline{E}_{xj}$, we can apply the same argument about the dependence on E_{yj} and conclude that we can have one of the three possible situations:

- $E_{yj} \in (\underline{E}_{yj}, \overline{E}_{yj})$ and $\dfrac{\partial \rho}{\partial E_{yj}} = 0$, i.e.,

$$E_{xj} = E_x + k_y \cdot (E_{yj} - E_y);$$

- $E_{yj} = \underline{E}_{yj}$ and $\dfrac{\partial \rho}{\partial E_{yj}} \le 0$, i.e.,

$$E_{xj} \le E_x + k_y \cdot (E_{yj} - E_y);$$

- $E_{yj} = \overline{E}_{yj}$ and $\dfrac{\partial \rho}{\partial E_{yj}} \ge 0$, i.e.,

$$E_{xj} \ge E_x + k_y \cdot (E_{yj} - E_y).$$

Here, k_y has the same sign as the correlation, so $k_y > 0$. Let us now consider possible locations of the segment $\{\overline{E}_{xj}\} \times [\underline{E}_{yj}, \overline{E}_{yj}]$ in relation to the y-line $E_{xj} = E_x + k_y \cdot (E_{yj} - E_y)$.

1.1°. The first subcase is when the whole segment is to the left of the y-line, i.e., when $E_{xj} < E_x + k_y \cdot (E_{yj} - E_y)$ for all $E_{yj} \in [\underline{E}_{yj}, \overline{E}_{yj}]$. In this case, we cannot have $E_{yj} \in (\underline{E}_{yj}, \overline{E}_{yj})$ and we cannot have $E_{yj} = \overline{E}_{yj}$, so we must have $E_{yj} = \underline{E}_{yj}$.

1.2°. The second subcase is when the whole segment is to the right of the y-line, i.e., when $E_{xj} > E_x + k_y \cdot (E_{yj} - E_y)$ for all $E_{yj} \in [\underline{E}_{yj}, \overline{E}_{yj}]$. In this case, we cannot have $E_{yj} \in (\underline{E}_{yj}, \overline{E}_{yj})$ and we cannot have $E_{yj} = \underline{E}_{yj}$, so we must have $E_{yj} = \overline{E}_{yj}$.

1.3°. The third subcase is when the segment intersects the y-line, i.e., when $E_{xj} = E_x + k_y \cdot (E'_{yj} - E_y)$ for some $E'_{yj} \in [\underline{E}_{yj}, \overline{E}_{yj}]$. As we have mentioned, there are three possibility for the value E_{yj} at which the correlation attains its maximum: the value for which $E_{xj} = E_x + k_y \cdot (E_{yj} - E_y)$, the value $\underline{E}_{yj}$, and the value $\overline{E}_{yj}$.

1.3.1°. In the first case (when $E_{xj} = E_x + k_y \cdot (E_{yj} - E_y)$), since $k_y > 0$, there is only one value $E_{yj} = E'_{yj}$.

1.3.2°. If $\underline{E}_{yj} \ne E'_{yj}$, then $\underline{E}_{yj} < E'_{yj}$, and thus,

$$E_x + k_y \cdot (\underline{E}_{yj} - E_y) < E_x + k_y \cdot (E'_{yj} - E_y) = E_{xj}.$$

Thus, we have $E_{xj} > E_x + k_y \cdot (\underline{E}_{yj} - E_y)$, so we cannot have $E_{xj} \le E_x + k_y \cdot (\underline{E}_{yj} - E_y)$, and therefore, the maximum cannot be attained for $E_{yj} = \underline{E}_{yj}$.

1.3.3°. If $\overline{E}_{yj} \ne E'_{yj}$, then $E'_{yj} < \overline{E}_{yj}$, and thus,

$$E_{xj} = E_x + k_y \cdot (E'_{yj} - E_y) < E_x + k_y \cdot (\overline{E}_{yj} - E_y) = E_{xj}.$$

Thus, we have $E_{xj} < E_x + k_y \cdot (\overline{E}_{yj} - E_y)$, so we cannot have $E_{xj} \le E_x + k_y \cdot (\overline{E}_{yj} - E_y)$, and therefore, maximum cannot be attained for $E_{yj} = \overline{E}_{yj}$.

1.3.4°. Therefore, in this third subcase, maximum can only be attained at the point on the y-line.

2°. The second case is when the whole box $[\underline{E}_{xj}, \overline{E}_{xj}] \times [\underline{E}_{yj}, \overline{E}_{yj}]$ is below the x-line $E_{yj} = E_y + k_x \cdot (E_{xj} - E_x)$, i.e., when $E_{yj} < E_y + k_x \cdot (E_{xj} - E_x)$ for all $E_{yj} \in [\underline{E}_{yj}, \overline{E}_{yj}]$ and $E_{xj} \in [\underline{E}_{xj}, \overline{E}_{xj}]$. In this case, we cannot have $E_{xj} \in (\underline{E}_{xj}, \overline{E}_{xj})$ and we cannot have $E_{xj} = \overline{E}_{xj}$, so we must have $E_{xj} = \underline{E}_{xj}$.

On the segment $E_{xj} = \underline{E}_{xj}$, we can apply the same argument about the dependence on E_{yj} as in Part 1 of this proof and come with the same conclusions.

3°. Same arguments apply if the whole box is fully to the left or to the right of the y-line. In this case, we have $E_{yj} = \overline{E}_{yj}$ or $E_{yj} = \underline{E}_{yj}$.

4°. The only remaining case is when the box intersects both with the x-line and with the y-line. In this case, similar to Part 1.3 of this proof, we conclude that the point (E_{xj}, E_{yj}) corresponding to the optimal tuple belongs both to the x-line and to the y-line. Thus, this point coincides with the intersection of these two lines.

In general, the x-line has the form $y - E_y = k_x \cdot (x - E_x)$. The y-line has the form $x - E_x = k_y \cdot (y - E_y)$, i.e., equivalently, $y - E_y = \dfrac{1}{k_y} \cdot (x - E_x)$. Both lines pass through the same point (E_x, E_y), but their slopes are, in general, different: k_x for the x-line and $\dfrac{1}{k_y}$ for the y-line. Thus, these lines coincide if and only if $k_x = \dfrac{1}{k_y}$, i.e., if and only if $k_x \cdot k_y = 1$.

In general, $\rho \le 1$. Here, $\rho = \dfrac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}$; thus, $\rho = \sqrt{k_x \cdot k_y}$, so $k_x \cdot k_y \le 1$. If $k_x \cdot k_y < 1$, then $k_x \cdot k_y \ne 1$ and thus, the x-line and the y-line are different. So, the intersection of these two lines is a single point (E_x, E_y). If $k_x \cdot k_y = 1$, this means that $\rho = 1$, and all the points (E_{xj}, E_{yj}) are on the same straight line – this is the case we have considered above.

ACKNOWLEDGMENT

The author is thankful to Vladik Kreinovich and to the anonymous referees for their guidance and suggestions.

REFERENCES

[1] A. Jalal-Kamali and V. Kreinovich, *Estimating Correlation under Interval Uncertainty*, University of Texas at El Paso, Department of Computer Science, Technical Report UTEP-CS-11-57, 2011, http://www.cs.utep.edu/vladik/tr11-57.pdf

[2] L. Longpré, G. Xiang, V. Kreinovich, and E. Freudenthal, "Interval Approach to Preserving Privacy in Statistical Databases: Related Challenges and Algorithms of Computational Statistics", In: V. Gorodetsky, I. Kotenko, and V. A. Skormin (eds.), *Proc. Int'l Conf. "Math. Methods, Models and Architectures for Computer Networks Security" MMM-ACNS-07*, St. Petersburg, Russia, September 13–15, 2007, Springer Lecture Notes in Computer Science, 2007, Vol. CCIS-1, pp. 346–361.

[3] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, New York, 2012.

[4] H. T. Nguyen and E. A. Walker, *A first course in fuzzy logic*, CRC Press, Boca Raton, Florida, 2005.

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

[6] G. Xiang and V. Kreinovich, "Estimating Variance under Interval and Fuzzy Uncertainty: Case of Hierarchical Estimation", In: P. Melin, O. Castillo, L. T. Aguilar, J. Kacprzyk, and W. Pedrycz (eds.), *Foundations of Fuzzy Logic and Soft Computing*, Proc. World Congress of the Int'l Fuzzy Systems Association IFSA'2007, Cancun, Mexico, June 18–21, 2007, Springer Lecture Notes on Artificial Intelligence, 2007, Vol. 4529, pp. 3–12.