Estimating Correlation under Interval Uncertainty

Ali Jalal-Kamali and Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
ajalalkamali@miners.utep.edu
vladik@utep.edu

Abstract

In many engineering 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. Measurements are never absolutely accurate; it is therefore necessary to take this inaccuracy into account when estimating the correlation ρ. 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 $[\widetilde{y}_i - \Delta_{yi}, \widetilde{y}_i + \Delta_{yi}]$. Different values from these intervals 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 $\underline{\rho} < 0$.

1 Introduction

Need for correlation. In engineering, we design systems for real-world applications. To make sure that the system functions correctly, we need to take into account all possible situations in which these systems will function. Each such situation can be characterized by the values of different quantities. To describe which combinations of these values are more probable and which are less probable, it is necessary 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

$$\rho = \frac{C}{\sigma_x \cdot \sigma_y},$$

of the covariance C to the product of standard deviations $\sigma_x = \sqrt{V_x}$ and $\sigma_y = \sqrt{V_y}$. 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 estimates as follows:

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

Known facts about correlation: brief reminder. It is known that the value of this correlation coefficient ρ is always between -1 and 1. The correlation is equal to 1 if and only if the values are positively linearly dependent, i.e., when for some coefficient $k_x > 0$, we have $y_i = E_y + k_x \cdot (x_i - E_x)$ for every i. The correlation is equal to -1 if and only if the values are negatively linearly dependent, i.e., when for some coefficient $k_x < 0$, we have $y_i = E_y + k_x \cdot (x_i - E_x)$ for every 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., [12]. 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:

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

The problem of computing the range of correlation under interval uncertainty is a particular case of the general problem of *interval computations* (see, e.g., [5, 9]): computing the range of a given function $f(x_1, \ldots, x_n)$ under the interval uncertainty $x_1 \in [\underline{x}_1, \overline{x}_1]$, $\ldots$, $x_n \in [\underline{x}_n, \overline{x}_n]$. Interval computations – in particular, interval computations of statistical characteristics – have many applications, in particular, engineering applications; see, e.g., [1, 4, 5, 7, 8, 9, 10, 11, 13].

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

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

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

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

Estimating correlation under interval uncertainty is NP-hard. In contrast to the mean, variance, covariance, and correlation are, in general, non-monotonic. It turns out that in general, computing the values of these characteristics under interval uncertainty is NP-hard [2, 3, 10, 11]. This means, crudely speaking, that unless P=NP (which most computable scientists believe to be wrong), no feasible (i.e., no polynomial-time) algorithm is possible that would always compute the range of the corresponding characteristic under interval uncertainty.

The problem of estimating correlation under interval uncertainty is formulated and analyzed in [13]; in that paper, this problem is formulated and solved as an optimization problem. For reasonably small n, the corresponding optimization algorithms work well [13]. However, since the problem is NP-hard, the computation time becomes infeasible when n is large.

What we do in this paper. We show 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. Specifically, we show 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 $0 \le \underline{0}$, 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 $\overline{\rho}$.

2 Main Result and the Corresponding Algorithm

Main result. *There exists a polynomial-time algorithm that, given n pairs of intervals $[\underline{x}_i, \overline{x}_i]$ and $[\underline{y}_i, \overline{y}_i]$, 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 $\overline{\rho} < 0$.*

Reducing minimum to maximum. When we change the sign of y_i, the correlation changes sign as well:

$$\rho(x_1, \ldots, x_n, -y_1, \ldots, -y_n) = -\rho(x_1, \ldots, x_n, y_1, \ldots, y_n).$$

Since the function $z \to -z$ is decreasing, its smallest value is attained when z is the largest, and its largest value is attained when z is the smallest. Thus, if z goes from $\underline{z}$ to $\overline{z}$, the range of $-z$ is $[-\overline{z}, -\underline{z}]$. So, for the endpoints of the ranges, we get

$$\overline{\rho}([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n], -[\underline{y}_1, \overline{y}_1], \ldots, -[\underline{y}_n, \overline{y}_n]) =$$

$$-\underline{\rho}([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n], [\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]),$$

where

$$-[\underline{y}_i, \overline{y}_i] = \{-y_i : y_i \in [\underline{y}_i, \overline{y}_i]\} = [-\overline{y}_i, -\underline{y}_i].$$

So, 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, as

$$\underline{\rho}([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n], [\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]) =$$

$$-\overline{\rho}([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n], [-\overline{y}_1, -\underline{y}_1], \ldots, [-\overline{y}_n, -\underline{y}_n]).$$

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

Algorithm. For each i from 1 to n, the corresponding box $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ has four vertices: $(\underline{x}_i, \underline{y}_i)$, $(\underline{x}_i, \overline{y}_i)$, $(\overline{x}_i, \underline{y}_i)$, and $(\overline{x}_i, \overline{y}_i)$. 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 x_i and y_i.

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 $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ 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 $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$, based on its location in comparison to the representative lines, we select the values x_i and y_i as follows:

- If the whole box is above the representative x-line, we take $x_i = \overline{x}_i$. On the resulting segment $\{\overline{x}_i\} \times [\underline{y}_i, \overline{y}_i]$, 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 $y_i = \underline{y}_i$;
  - if the whole segment is to left of the representative y-line, we select $y_i = \overline{y}_i$;
  - if the segment intersects with the representative y-line, we select the value y_i 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 $x_i = \underline{x}_i$. On the resulting segment $\{\underline{x}_i\} \times [\underline{y}_i, \overline{y}_i]$, 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 $y_i = \underline{y}_i$;
  - if the whole segment is to left of the representative y-line, we select $y_i = \overline{y}_i$;
  - if the segment intersects with the representative y-line, we select the value y_i 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 $y_i = \underline{y}_i$. On the resulting segment $[\underline{x}_i, \overline{x}_i] \times \{\underline{y}_i\}$, 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 $x_i = \underline{x}_i$;
  - if the whole segment is below the representative x-line, we select $x_i = \overline{x}_i$;
  - if the segment intersects with the representative x-line, we select the value x_i 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 $y_i = \overline{y}_i$. On the resulting segment $[\underline{x}_i, \overline{x}_i] \times \{\overline{y}_i\}$, 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 $x_i = \underline{x}_i$;
  - if the whole segment is below the representative x-line, we select $x_i = \overline{x}_i$;
  - if the segment intersects with the representative x-line, we select the value x_i 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 i and for each of the values x_i and y_i, 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 x_i and y_i into the following formulas, we get a system of four equations with four unknowns E_x, E_y, k_x and k_y:

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

$$\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - E_x \cdot E_y = k_x \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x)^2 \right);$$

$$\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - E_x \cdot E_y = k_y \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - E_y)^2 \right).$$

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 x_i and y_i.

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 i, according to our algorithm, as the appropriate value of x_i, we make one of the following four selections:

- sometimes, we select a known value $\underline{x_i}$;
- sometimes, we select a know value $\overline{x}_i$;
- sometimes, we select the value $x_i = E_x$ (which is not a priori known, it is one of the four variables that we need to determine), and
- sometimes, we select a value x_i that lies on the x-line $y = E_y + k_x \cdot (x_i - E_x)$, i.e., a value $x_i = E_x + K_x \cdot (y_i - E_y)$, where $K_x \stackrel{\text{def}}{=} \dfrac{1}{k_x} = \dfrac{V_x}{C}$.

In general, each expression x_i 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 x_i.

Similarly, for each i, as the appropriate value of y_i, we make one of the following four selections:

- sometimes, we select a known value $\underline{y}_i$;
- sometimes, we select a know value $\overline{y}_i$;
- sometimes, we select the value $y_i = E_y$ (which is not a priori known, it is one of the four variables that we need to determine), and
- sometimes, we select a value y_i that lies on the y-line $x = E_x + k_y \cdot (y_i - E_y)$, i.e., a value $y_i = E_y + K_y \cdot (x_i - E_x)$, where $K_y \stackrel{\text{def}}{=} \dfrac{1}{k_y} = \dfrac{V_y}{C}$.

In general, each expression y_i is a linear combination of a constant and the unknowns E_y, K_y, and $K_y \cdot E_x$.

Substituting these expressions for x_i and y_i into the four equations for the unknowns E_x, E_y, K_x, and K_y, we conclude that:

- the equation $E_x = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$ is transformed into equating a linear combination of E_x, K_x, and $K_x \cdot E_y$, to zero;
- the equation $E_y = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} y_i$ is transformed into equating a linear combination of E_y, K_y, and $K_y \cdot E_x$, to zero;
- the equation $V_x = K_x \cdot C$, i.e.,

  $$K_x \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - E_x \cdot E_y \right) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x)^2$$

  is transformed into equating a linear combination of terms of order ≤ 4 in terms of the unknowns;
- we also get a similar transformation for the equation $V_y \cdot K_y \cdot C$.

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 n of the original sample, so in terms of dependence on this size, we need $O(1)$ time.

3 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 $4n$ possible vertices, so we have $O(n^2)$ possible pairs of vertices – and thus, $O(n^2)$ possible 4-tuples. Thus, we have $O(n^2)$ possible representative x-lines, and we also have $O(n^2)$ representative y-lines. In our algorithms,

we consider pairs consisting of a representative x-line and a representative y-line. Since we have $O(n^2)$ x-lines and we have $O(n^2)$ y-lines, we therefore have $O(n^2) \cdot O(n^2) = O(n^4)$ 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 n values x_i and each of n values y_i in terms of the parameters E_x, E_y, K_x, and K_y. So, we need $O(n)$ steps to find these expressions for all i.
- Then, we need linear time $O(n)$ 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 $O(n)$ to check whether each of $4n = O(n)$ vertices is in the right position with respect to the corresponding lines, and,
  - if needed, linear time $O(n)$ to compute the corresponding value of the correlation ρ – by using the above explicit formula describing how the correlation ρ depends on x_i and y_i.

Totally, for each pair of lines, we need

$$O(n) + O(n) + O(1) + O(n) + O(n) = O(n)$$

computational steps.

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

Case when the representative x-line coincides with the representative y-line. If this common line intersects with all n boxes $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$, then, for each box, we can select values x_i and y_i for which the corresponding point (x_i, y_i) belongs to this line. Then, all selected values (x_i, y_i) follow the same linear dependence $y_i = E_y + k_x \cdot (x_i - E_x)$ (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: $\frac{df}{dx} = 0$.

If it attains its maximum at the point $x = \underline{x}$, then we cannot have $\frac{df}{dx} > 0$, 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 $\frac{df}{dx} \le 0$.

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

Computing the corresponding derivatives. We are interested in the values x_i and y_i for which the correlation ρ attains maximum. To use the above facts, let us find the partial derivatives of ρ with respect to x_i and y_i.

The correlation is defined as the ratio of the covariance C and the product of the standard deviations σ_x and σ_y. These quantities, in their turn, are described in terms of V_x, V_y, E_x, and E_y. To compute the corresponding partial derivative, let us first compute the partial derivatives of E_x and E_y, then of V_x, V_y, and C, and then finally, of the correlation ρ.

Based on the above expression for E_x, we conclude that $\frac{\partial E_x}{\partial x_i} = \frac{1}{n}$ and similarly $\frac{\partial E_y}{\partial y_i} = \frac{1}{n}$. Since the variance V_x can be described in an equivalent form $V_x = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - E_x^2$, we get

$$\frac{\partial V_x}{\partial x_i} = \frac{2}{n} \cdot x_i - 2 \cdot E_x \cdot \frac{\partial E_x}{\partial x_i} = \frac{2}{n} \cdot (x_i - E_x).$$

Similarly,

$$\frac{\partial V_y}{\partial y_i} = \frac{2}{n} \cdot (y_i - E_y).$$

Now, since $\sigma_x = \sqrt{V_x}$, we have

$$\frac{\partial \sigma_x}{\partial x_i} = \frac{1}{2} \cdot \frac{1}{\sqrt{V_x}} \cdot \frac{\partial V_x}{\partial x_i} = \frac{1}{2} \cdot \frac{1}{\sigma_x} \cdot \frac{\partial V_x}{\partial x_i}.$$

Substituting the above formula for the derivative of V_x, we get $\frac{\partial \sigma_x}{\partial x_i} = \frac{x_i - E_x}{n \cdot \sigma_x}$ and similarly, $\frac{\partial \sigma_y}{\partial y_i} = \frac{y_i - E_y}{n \cdot \sigma_y}$.

Now, since $C = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - E_x \cdot E_y$, we get

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

Thus, for $\rho = \dfrac{C}{\sigma_x \cdot \sigma_y}$, since σ_y does not depend on x_i, we get

$$\frac{\partial \rho}{\partial x_i} = \frac{1}{\sigma_y} \cdot \frac{\partial}{\partial x_i}\left(\frac{C}{\sigma_x}\right) = \frac{1}{\sigma_y} \cdot \frac{\dfrac{\partial C}{\partial x_i} \cdot \sigma_x - C \cdot \dfrac{\partial \sigma_x}{\partial x_i}}{\sigma_x^2} =$$

$$\frac{1}{\sigma_y \cdot \sigma_x^2 \cdot n} \cdot \left[(y_i - E_y) \cdot \sigma_x - C \cdot \frac{x_i - E_x}{\sigma_x}\right].$$

Since the standard deviations are always non-negative, the sign of this derivative coincides with the sign of the value $(y_i - E_y) \cdot \sigma_x - C \cdot \dfrac{x_i - E_x}{\sigma_x}$. Dividing this expression by a positive value σ_x, we conclude that the sign of the derivative $\dfrac{\partial \rho}{\partial x_i}$ coincides with the sign of the expression $(y_i - E_y) - k_x \cdot (x_i - E_x)$, where we denoted $k_x \stackrel{\text{def}}{=} \dfrac{C}{V_x}$.

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

It is worth mentioning since the standard deviations and variances are non-negative, the sign of both coefficients $k_x = \dfrac{C}{V_x}$ and $k_y = \dfrac{C}{V_x}$ coincides with the sign of the correlation $\rho = \dfrac{C}{\sigma_x \cdot \sigma_y}$.

Let us apply the known facts from calculus to this situation. Let x_i and y_i 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:

- $x_i \in (\underline{x}_i, \overline{x}_i)$ and $\dfrac{\partial \rho}{\partial x_i} = 0$, i.e., $y_i = E_y + k_x \cdot (x_i - E_x)$;
- $x_i = \underline{x}_i$ and $\dfrac{\partial \rho}{\partial x_i} \le 0$, i.e., $y_i \le E_y + k_x \cdot (x_i - E_x)$;
- $x_i = \overline{x}_i$ and $\dfrac{\partial \rho}{\partial x_i} \ge 0$, i.e., $y_i \ge E_y + k_x \cdot (x_i - E_x)$.

Here, k_x has the same sign as the correlation, so $k_x > 0$. Let us now consider possible locations of the box $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ with respect to the x-line $y_i = E_y + k_x \cdot (x_i - E_x)$.

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

On the segment $x_i = \overline{x}_i$, we can apply the same argument about the dependence on y_i and conclude that we can have one of the three possible situations:

- $y_i \in (\underline{y}_i, \overline{y}_i)$ and $\dfrac{\partial \rho}{\partial y_i} = 0$, i.e., $x_i = E_x + k_y \cdot (y_i - E_y)$;
- $y_i = \underline{y}_i$ and $\dfrac{\partial \rho}{\partial y_i} \le 0$, i.e., $x_i \le E_x + k_y \cdot (y_i - E_y)$;
- $y_i = \overline{y}_i$ and $\dfrac{\partial \rho}{\partial y_i} \ge 0$, i.e., $x_i \ge E_x + k_y \cdot (y_i - E_y)$.

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

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

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

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

1.3.1°. In the first case (when $x_i = E_x + k_y \cdot (y_i - E_y)$), since $k_y > 0$, there is only one value $y_i = y_i'$.

1.3.2°. If $\underline{y}_i \ne y_i'$, then $\underline{y}_i < y_i'$, and thus,

$$E_x + k_y \cdot (\underline{y}_i - E_y) < E_x + k_y \cdot (y_i' - E_y) = x_i.$$

Thus, we have $x_i > E_x + k_y \cdot (\underline{y}_i - E_y)$, so we cannot have $x_i \le E_x + k_y \cdot (\underline{y}_i - E_y)$, and therefore, the maximum cannot be attained for $y_i = \underline{y}_i$.

1.3.3°. If $\overline{y}_i \ne y_i'$, then $y_i' < \overline{y}_i$, and thus,

$$x_i = E_x + k_y \cdot (y_i' - E_y) < E_x + k_y \cdot (\overline{y}_i - E_y) = x_i.$$

Thus, we have $x_i < E_x + k_y \cdot (\overline{y}_i - E_y)$, so we cannot have $x_i \le E_x + k_y \cdot (\overline{y}_i - E_y)$, and therefore, maximum cannot be attained for $y_i = \overline{y}_i$.

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 $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ is below the x-line $y_i = E_y + k_x \cdot (x_i - E_x)$, i.e., when $y_i < E_y + k_x \cdot (x_i - E_x)$ for all $y_i \in [\underline{y}_i, \overline{y}_i]$

and $x_i \in [\underline{x}_i, \overline{x}_i]$. In this case, we cannot have $x_i \in (\underline{x}_i, \overline{x}_i)$ and we cannot have $x_i = \overline{x}_i$, so we must have $x_i = \underline{x}_i$.

On the segment $x_i = \underline{x}_i$, we can apply the same argument about the dependence on y_i 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 $y_i = \overline{y}_i$ or $y_i = \underline{y}_i$.

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 (x_i, y_i) 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, $\rho = \dfrac{C}{\sigma_x \cdot \sigma_y} = \dfrac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}$; 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 (x_i, y_i) are on the same straight line – this is the case we have considered above.

5°. We enumerated all the cases described in the algorithm and showed that in all these cases, we should produce exactly the values x_i and y_i described in the algorithm. Thus, we have justified the algorithm – provided that we enumerate all possible locations of the vertices with respect to x- and y-lines.

To complete the proof, we need to show that all possible locations are captured by what we called representative x- and y-lines. Indeed, let us start with any x-line, and let us show that there exists a representative x-line that has exactly the same location with respect to all the vertices – i.e., that each vertex is above, below, or on the representative x-line if and only if this vertex is, correspondingly, above, below, or on the actual x-line.

Let us take the actual x-line. It is contains one of the vertices, mark this vertex. If the original x-line does not contain any of the vertices, let us move the line (parallel to itself) along the x-axis – until the line hits a vertex. Then, we move the line back by a small amount, and we mark this almost-vertex point.

Once the marked vertex is fixed, we check if the line contains another vertex. If it does, we mark that vertex, and so we have the desired representative x-line. If it does not, we rotate the line around the already marked vertex (or almost-vertex) until the line starts containing another vertex. We similarly move the

line back by a small amount, and we get the desired representative x-line that is in exactly same relation to all the vertices as the actual x-line.

We can perform the same procedure with the y-line. Correctness is proven.

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

References

[1] C. Ferregut, F. J. Campos, and V. Kreinovich, "Reducing over-conservative expert failure rate estimates in the presence of limited data: a new probabilistic/fuzzy approach", *Proceedings of the 30th Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2011*, El Paso, Texas, March 18–20, 2011.

[2] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Computing Variance for Interval Data is NP-Hard", *ACM SIGACT News*, vol. 33, no. 2, pp. 108–118, 2002.

[3] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Exact Bounds on Finite Populations of Interval Data", *Reliable Computing*, vol. 11, no. 3, pp. 207–233, 2005.

[4] C. Jacob, D. Dubois, J. Cardoso, M. Ceberio, and V. Kreinovich, "Estimating Probability of Failure of a Complex System Based on Partial Information about Subsystems and Components, with Potential Applications to Aircraft Maintenance", *Proceedings of the International Workshop on Soft Computing Applications and Knowledge Discovery SCAKD'2011*, Moscow, Russia, June 25, 2011, pp. 30–41.

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

[6] V. Kreinovich, "Reliability Analysis for Aerospace Applications: Reducing Over-Conservative Expert Estimates in the Presence of Limited Data", In: Sergei O. Kuznetsov and Dominik Slezak (eds.), *Expert and Industry Sessions of the 13th International Conference on Rough Sets, Fuzzy Sets and Granular Computing RSFDGrC'2011 and the 4th International Conference on Pattern Recognition and Machine Intelligence PReMI'2011*, Moscow, Russia, June 25–30, 2011.

[7] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos, "Towards combining probabilistic and interval uncertainty in engineering calculations: al-

gorithms for computing statistics under interval uncertainty, and their computational complexity", *Reliable Computing*, vol. 12, no. 6, pp. 471–501, 2006.

[8] V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos, "Interval Versions of Statistical Techniques, with Applications to Environmental Analysis, Bioinformatics, and Privacy in Statistical Databases", *Journal of Computational and Applied Mathematics*, vol. 199, no. 2, pp. 418–423, 2007.

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

[10] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, to appear.

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

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

[13] F. Tonon and C. L. Pettit, "Toward a definition and understanding of correlation for variables constrained by random relations", *International Journal of General Systems*, 2010, Vol. 39, No. 6, pp. 577–604.