

ESTIMATING STATISTICAL CHARACTERISTICS UNDER  
INTERVAL UNCERTAINTY AND CONSTRAINTS:  
MEAN, VARIANCE, COVARIANCE, AND CORRELATION

ALI JALAL-KAMALI

Department of Computer Science

APPROVED:

---

Vladik Kreinovich, Chair, Ph.D.

---

Luc Longpré, Ph.D.

---

Peter Moschopoulos, Ph.D.

---

Benjamin C. Flores, Ph.D.  
Dean of the Graduate School

©Copyright

by

Ali Jalal-Kamali

2011

*to my*

*FATHER and MOTHER and ALL OTHER MEMBERS OF MY FAMILY*

*with love*

ESTIMATING STATISTICAL CHARACTERISTICS UNDER  
INTERVAL UNCERTAINTY AND CONSTRAINTS:  
MEAN, VARIANCE, COVARIANCE, AND CORRELATION

by

ALI JALAL-KAMALI

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

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

December 2011

# Acknowledgements

I would like to express my high respect and regards to my advisor, Dr. Vladik Kreinovich, for his wise advices, and constant support. For me, he has been the best professor I have ever had, and no matter what time of the day, whenever I showed up with a question at his office, I always got help, despite his busy schedule. His kind support was not limited to school only. He has invited me to other places and events, and also every now and then he would have a conversation with me about my country, so that I don't feel too lonely.

I also wish to thank the other members of my committee, Dr. Luc Longpré and Dr. Peter Moschopoulos. Dr. Longpré, also like Dr. Kreinovich, has been extremely supportive and helpful since I came to El Paso both in terms of school related and non-related issues. He helped me a lot to deal with many issues all through my education, and to make me feel like home he has introduced and invited me to many interesting activities in the city.

Additionally, I want to thank all professors and staff from the University of Texas at El Paso Computer Science Department for all their hard work and dedication, providing me the means to complete my degree and prepare for a career as a computer scientist.

In particular, I have to express my deep gratitude to Dr. Eric Freudenthal. His kind support started even before I come to to USA, by him introducing places to rent in El Paso, and it continued since then. He helped me to get to know the points of interest in the city, specially the ones that were Persian related. He also added me to his engineering education research team, and suggested me as the instructor for CS1420. I have walked to his office with issues ranging from very personal all the way to scientific questions, and regardless of the origin of my problem, he always has helped me without any hesitations.

I am pleased to say that besides my family, and my close friends, Dr. Kreinovich, Dr. Freudenthal, and Dr. Longpré are among the people for whom I have a very high respect in my heart.

And finally, I must thank my amazing family for their non-stop love and support all

through my life. Their love and support got me to this point enabling me to go through everything to be successful. There are no words to express all my feelings, I can only say THANK YOU for everything!

NOTE: This thesis was submitted to my Supervising Committee on the November 20, 2011.

# Abstract

In many practical situations, we have a sample of objects of a given type. When we measure the values of a certain quantity  $x$  for these objects, we get a sequence of values  $x_1, \dots, x_n$ . When the sample is large enough, then the arithmetic mean  $E$  of the values  $x_i$  is a good approximation for the average value of this quantity for all the objects from this class. Other expressions provide a good approximation to statistical characteristics such as variance, covariance, and correlation.

The values  $x_i$  come from measurements, and measurement is never absolutely accurate. Often, the only information that we have about the measurement error is the upper bound  $\Delta_i$  on this error. In this case, once we have the measurement result  $\tilde{x}_i$ , the condition  $|\tilde{x}_i - x_i| \leq \Delta_i$  implies that the actual (unknown) value  $x_i$  belongs to the interval  $\mathbf{x}_i \stackrel{\text{def}}{=} [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$ . Different values  $x_i \in \mathbf{x}_i$  from the corresponding intervals lead, in general, to different values of sample mean, sample variance, etc. It is therefore desirable to find the range of possible values of these characteristics when  $x_i \in \mathbf{x}_i$ .

It is known that evaluating such ranges is, in general, NP-hard (see, e.g., [2]). The main objective of this thesis is to design feasible (i.e., polynomial-time) algorithms for practically important situations. Several such algorithms are described and proved to be correct.

# Table of Contents

	<b>Page</b>
Acknowledgements . . . . .	v
Abstract . . . . .	vii
Table of Contents . . . . .	viii
<b>Chapter</b>	
1 Introduction . . . . .	1
1.1 Need for Estimating Statistical Characteristics . . . . .	1
1.2 Case of Interval Uncertainty . . . . .	2
1.3 Need to Preserve Privacy in Statistical Databases . . . . .	2
1.4 Intervals As a Way to Preserve Privacy in Statistical Databases . . . . .	3
1.5 Need to Estimate Statistical Characteristics Under Interval Uncertainty . .	4
1.6 Estimating Statistical Characteristics Under Interval Uncertainty: What is Known . . . . .	5
1.7 Our Results . . . . .	5
2 Estimating Mean under Interval Uncertainty and Variance Constraint . . . . .	7
2.1 Formulation of the Problem . . . . .	7
2.2 Estimating Mean Under Interval Uncertainty and Variance Constraint: a Problem . . . . .	8
2.3 Main Result . . . . .	10
2.4 Computation Time of This Algorithm . . . . .	11
2.5 Simple Example . . . . .	12
2.6 Proof of the Algorithm's Correctness . . . . .	14
3 Estimating Covariance for the Privacy Case under Interval Uncertainty . . . . .	21
3.1 Formulation of the Problem . . . . .	21
3.2 Analysis of the Problem . . . . .	22

3.3	Resulting Algorithm . . . . .	27
3.4	Computation Time of the Algorithm . . . . .	32
4	Estimating Correlation under Interval Uncertainty . . . . .	33
4.1	Introduction . . . . .	33
4.2	Main Result and the Corresponding Algorithm . . . . .	34
4.3	Proof of the Main Result . . . . .	40
5	Concluding Remarks . . . . .	47
	References . . . . .	49
	Curriculum Vitae . . . . .	52

# Chapter 1

## Introduction

### 1.1 Need for Estimating Statistical Characteristics

In many practical situations, we have a sample of values  $x_1, \dots, x_n$  corresponding to objects of a certain type. For example,  $x_i$  may represent the height of the  $i$ -th person in a group, or his or her weight, or the toxicity of the  $i$ -th snake of a certain species.

In this case, a standard way to describe the corresponding population is to estimate its mean

$$E = \frac{1}{n} \cdot \sum_{i=1}^n x_i \quad (1.1)$$

and its variance

$$V = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - E)^2. \quad (1.2)$$

and standard deviation  $\sigma = \sqrt{V}$ .

In situations when we measure two quantities  $x_i$  and  $y_i$  for each object  $i$ , then we would like to describe the mean, variance, and standard deviation of each of these characteristics

$$\begin{aligned} E_x &= \frac{1}{n} \cdot \sum_{i=1}^n x_i, & E_y &= \frac{1}{n} \cdot \sum_{i=1}^n y_i, \\ V_x &= \frac{1}{n} \cdot \sum_{i=1}^n (x_i - E_x)^2, & V_y &= \frac{1}{n} \cdot \sum_{i=1}^n (y_i - E_y)^2, \\ \sigma_x &= \sqrt{V_x}, & \sigma_y &= \sqrt{V_y}, \end{aligned}$$

and we would also like to know their covariance

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

and their correlation

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

## 1.2 Case of Interval Uncertainty

The above formulas implicitly assume that we know the exact values of the characteristics  $x_1, \dots, x_n$ . In practice, these values usually come from measurements, and measurements are never absolutely exact (see, e.g., [15]): the measurement results  $\tilde{x}_i$  are, in general, different from the actual (unknown) values  $x_i$ :  $\tilde{x}_i \neq x_i$ .

In the traditional engineering and scientific practice, it is usually assumed that we know the probability distribution of the measurement errors  $\Delta x_i \stackrel{\text{def}}{=} \tilde{x}_i - x_i$ . However, often, the only information we have is the upper bound  $\Delta_i$  on the (absolute value of the) measurement error:  $|\Delta x_i| \leq \Delta_i$ ; see, e.g., [15].

In this case, based on the measurement result  $\tilde{x}_i$ , the only information that we have about the actual (unknown) value  $x_i$  is that  $x_i$  belongs to the interval  $\mathbf{x}_i = [\underline{x}_i, \bar{x}_i]$ , where  $\underline{x}_i = \tilde{x}_i - \Delta_i$  and  $\bar{x}_i = \tilde{x}_i + \Delta_i$ .

## 1.3 Need to Preserve Privacy in Statistical Databases

In addition to measurement errors, there is another important source of interval uncertainty: the need to preserve privacy in statistical databases. In order to find relations between different quantities, we collect a large amount of data. For example, we collect a large amount of medical data – to try to find relation between instances of a certain disease and lifestyle factors that may contribute to this disease. We collect a large amount of data in a census – to see, e.g., how the parents’ income level affects the children’s education level, and how the person education level influences his or her income level.

In some cases, we are looking for commonsense associations – e.g., between smoking and lung diseases, obesity and diabetes, etc. However, in many cases, it is not clear which

factors affect a certain disease. For example, if a rare disease appears in certain areas, it may be because of the high concentration of some chemical in these areas, but we often do not know *a priori* which chemicals to look for.

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

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

It is therefore desirable to make sure that privacy is preserved in statistical databases.

## 1.4 Intervals As a Way to Preserve Privacy in Statistical Databases

One way to preserve privacy is not to store the exact data values – from which a person can be identified – in the database, but rather store *ranges* (intervals).

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

And this is how data is often collected: instead of asking for an exact age, we ask a

person to check whether her age is, say, in between 0 and 10, 10 and 20, etc. Similarly, instead of the exact income, we ask the person to indicate into which income bracket his or her income falls.

In general, we set some threshold values  $t_0, \dots, t_N$  and ask a person whether the actual value of the corresponding quantity is in the interval  $[t_0, t_1]$ , in the interval  $[t_1, t_2]$ ,  $\dots$ , or in the interval  $[t_{N-1}, t_N]$ .

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

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

## 1.5 Need to Estimate Statistical Characteristics Under Interval Uncertainty

In both situations of measurement errors or privacy, instead of the actual values  $x_i$  (and  $y_i$ ), we only know the intervals  $\mathbf{x}_i$  (and  $\mathbf{y}_i$ ) that contain the actual (unknown) values. Different values of  $x_i$  (and  $y_i$ ) from these intervals lead, in general, to different values of each statistical characteristic  $S(x_1, \dots, x_n)$  (or  $S(x_1, \dots, x_n, y_1, \dots, y_n)$ ). It is therefore desirable to find the *range* of possible values of these characteristics when  $x_i \in \mathbf{x}_i$  (and  $y_i \in \mathbf{y}_i$ ):

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

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

## 1.6 Estimating Statistical Characteristics Under Interval Uncertainty: What is Known

The general problem of estimating the range of a function under interval uncertainty is known as *interval computations* (see, e.g., [7, 12]): when the only information that we have about the actual value  $x_i$  is that this value is in the interval  $\mathbf{x}_i = [\underline{x}_i, \bar{x}_i]$ , then, for each auxiliary quantity  $y = f(x_1, \dots, x_n)$ , we get an interval of possible values of  $y$ :

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

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  $\bar{E}_x$  is attained when each of the variables attains its largest value  $\bar{x}_i$ :

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

However, variance, covariance, and correlation are, in general, non-monotonic. It turns out that in general, computing the values of these characteristics under interval uncertainty is NP-hard [2, 3, 14]. This means, crudely speaking, that unless P=NP (which most computable scientists believe to be wrong), no feasible (polynomial-time) algorithm is possible that would always compute the range of the corresponding characteristic under interval uncertainty.

The problem gets even more complex if we take into account that in practice, we often have additional constraints on possible combinations of the values  $x_i$  – and when estimating the range, we need to only consider the values that satisfy these additional constraints.

## 1.7 Our Results

Usually, even when a general problem is NP-hard, there are practically important cases when a feasible algorithm is possible. The main objective of this thesis is to find such cases

for estimating statistical characteristics under interval uncertainty.

We start with the simplest of the statistical characteristics – the mean. In the absence of constraints, as we have mentioned, computing the range of the mean under interval uncertainty is easy. However, the presence of constraints makes the problem much more difficult. In Chapter 2, we describe a feasible algorithm for estimating the mean under interval uncertainty in the presence of constraints limiting possible values of sample variance.

Other characteristics of interest are variance, covariance, and correlation. As we have mentioned, in general, for all three characteristics, the problem of estimating the corresponding ranges is NP-hard, so it is desirable to find classes of problems for which feasible algorithms are possible. For variance, a lot of feasible algorithms have already been designed [2, 3, 13, 17]. In particular, it is known that computing one of the endpoints for the range of variance is always feasible, and that in the privacy case, we can feasibly compute both endpoints. In this thesis, we extend these results to covariance and correlation. Specifically, in Chapter 3, we analyze covariance, and we show that for the case when interval uncertainty comes from privacy, a feasible algorithm is possible.

In Chapter 4, we analyze correlation, and we show that while it is not possible to feasible compute both endpoints  $\underline{\rho}$  and  $\bar{\rho}$  of the corresponding interval range  $[\underline{\rho}, \bar{\rho}]$ , we can always compute at least one of the endpoints: namely, we can feasibly compute the upper endpoint  $\bar{\rho}$  when it is positive, and we can feasibly compute the lower endpoint  $\underline{\rho}$  when it is negative.

# Chapter 2

## Estimating Mean under Interval Uncertainty and Variance Constraint

In this chapter we show how to have a feasible algorithm for estimating mean under interval uncertainty and variance constraint. The results of this chapter were published in [8] and [6].

### 2.1 Formulation of the Problem

As we have mentioned in Chapter 1, estimating the range of the mean  $E$  under interval uncertainty is an easy computational problem, while the problem of computing the range  $[\underline{V}, \bar{V}]$  of the variance  $V$  under interval uncertainty is, in general, NP-hard. Specifically, it turns out that while the lower endpoint  $\underline{V}$  can be computed in linear time [13, 17, 18], the problem of computing  $\bar{V}$  is NP-hard [2, 3, 13, 17].

In these papers, it was assumed that there is no a priori information about the values of  $E$  and  $V$ . In some cases, we have *a priori* constraints on the variance:  $V \leq V_0$  for a given  $V_0$ . For example, we know that within a species, there can be no more than 0.1 variation of a certain characteristic. In such cases, we face a problem of estimating the values of the corresponding statistical characteristics under interval uncertainty *and* the variance constraint.

## 2.2 Estimating Mean Under Interval Uncertainty and Variance Constraint: a Problem

**Formulation of the problem.** In the presence of variance constraints, the problem of finding possible values of the mean  $E$  takes the following form:

- *given:*  $n$  intervals  $\mathbf{x}_i = [\underline{x}_i, \bar{x}_i]$  and a number  $V_0 \geq 0$ ;
- *compute:* the range

$$[\underline{E}, \bar{E}] = \{E(x_1, \dots, x_n) \mid x_i \in \mathbf{x}_i \ \& \ V(x_1, \dots, x_n) \leq V_0\}; \quad (2.1)$$

- *under the assumption* that there exist values  $x_i \in \mathbf{x}_i$  for which  $V(x_1, \dots, x_n) \leq V_0$ .

This is a problem that we will solve in this paper.

**Special case where the solution is straightforward.** Let us first consider the case when  $V_0$  is larger than (or equal to) the largest possible value  $\bar{V}$  of the variance corresponding to the given sample.

In this case, the constraint  $V \leq V_0$  is always satisfied. Thus, in this case, the desired range simply coincides with the range of all possible values of  $E$ :

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

*Comment.* It should be mentioned that the computation of the range  $[\underline{E}, \bar{E}]$  is easy only if we already *know* that  $\bar{V} \leq V_0$ .

Checking whether this inequality is satisfied is, as we have mentioned, a computationally difficult (NP-hard) problem; see, e.g., [2, 3].

**Special case when this problem is (relatively) easy to solve.** Another such case is when  $V_0 = 0$ .

In this case, the constraint  $V \leq V_0$  means that the variance  $V$  should be equal to 0. In this case, all non-negative values  $(x_i - E)^2$  should also be equal to 0 – otherwise, the average  $V$  of these values  $(x_i - E)^2$  would be positive. So, we have  $x_i = E$  for all  $i$  and thus, all the actual (unknown) values should coincide:  $x_1 = \dots = x_n$ . In this case, we know that this common value  $x_i$  belongs to each of  $n$  intervals  $\mathbf{x}_i$ , so it belongs to their *intersection*.

$$\mathbf{x}_1 \cap \dots \cap \mathbf{x}_n. \quad (2.2)$$

A value  $E$  belongs to the interval  $[\underline{x}_i, \bar{x}_i]$  if it is larger than or equal to its lower endpoint  $\underline{x}_i$  and smaller than or equal to its upper endpoint  $\bar{x}_i$ . Thus, for a value  $E$  to belong to all  $n$  intervals, it has to be larger than or equal to all  $n$  lower endpoints  $\underline{x}_1, \dots, \underline{x}_n$ , and it has to be smaller than or equal to all  $n$  upper endpoints  $\bar{x}_1, \dots, \bar{x}_n$ .

A number  $E$  is larger than or equal to  $n$  given numbers  $\underline{x}_1, \dots, \underline{x}_n$  if and only if it is larger than or equal to the largest of these  $n$  numbers, i.e., if  $\max(\underline{x}_1, \dots, \underline{x}_n) \leq E$ . Similarly, a number  $E$  is smaller than or equal to  $n$  given numbers  $\bar{x}_1, \dots, \bar{x}_n$  if and only if it is smaller than or equal to the smallest of these  $n$  numbers, i.e., if  $E \leq \min(\bar{x}_1, \dots, \bar{x}_n)$ . So, the intersection consists of all the numbers which are located between these two bounds, i.e., the intersection coincides with the interval

$$[\underline{E}, \bar{E}] = [\max(\underline{x}_1, \dots, \underline{x}_n), \min(\bar{x}_1, \dots, \bar{x}_n)]. \quad (2.3)$$

*Comment.* In this case, not only computing the range is easy, it is also easy to check whether there exist values  $x_i \in \mathbf{x}_i$  for which  $V(x_1, \dots, x_n) \leq V_0 = 0$ .

Indeed, as we have mentioned, this inequality is equivalent to the fact that  $x_1 = \dots = x_n$ . Thus, there exist values  $x_i \in \mathbf{x}_i$  that satisfy this inequality if and only if  $n$  intervals  $\mathbf{x}_i = [\underline{x}_i, \bar{x}_i]$  have a common element, i.e., if and only if

$$\max(\underline{x}_1, \dots, \underline{x}_n) \leq \min(\bar{x}_1, \dots, \bar{x}_n).$$

**General case.** In the general case, when  $V_0$  is larger than 0 but smaller than the upper endpoint  $\bar{V}$ , we should get intervals intermediate between intersection and arithmetic average. In this paper, we show how to compute the corresponding interval for  $E$ .

## 2.3 Main Result

**Algorithm.** The following feasible algorithm solves the problem of computing the range  $[\underline{E}, \bar{E}]$  of the mean under interval uncertainty and variance constraint:

- First, we compute the values

$$E^- \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^n \underline{x}_i \text{ and } V^- \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^n (\underline{x}_i - E^-)^2;$$

$$E^+ \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^n \bar{x}_i \text{ and } V^+ \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^n (\bar{x}_i - E^+)^2.$$

- If  $V^- \leq V_0$ , then we return  $\underline{E} = E^-$ .
- If  $V^+ \leq V_0$ , then we return  $\bar{E} = E^+$ .
- If at least one of these inequalities does not hold, i.e., if  $V_0 < V^-$  or  $V_0 < V^+$ , then we sort all  $2n$  endpoints  $\underline{x}_i$  and  $\bar{x}_i$  into a non-decreasing sequence

$$z_1 \leq z_2 \leq \dots \leq z_{2n}$$

and consider  $2n - 1$  zones  $[z_k, z_{k+1}]$ .

- For each zone  $[z_k, z_{k+1}]$ , we take:
  - for every  $i$  for which  $\bar{x}_i \leq z_k$ , we take  $x_i = \bar{x}_i$ ;
  - for every  $i$  for which  $z_{k+1} \leq \underline{x}_i$ , we take  $x_i = \underline{x}_i$ ;
  - for every other  $i$ , we take  $x_i = \alpha$ ; let us denote the number of such  $i$ 's by  $n_k$ .

The value  $\alpha$  is determined from the condition that for the selected vector  $x$ , we have  $V(x) = V_0$ , i.e., from solving the following quadratic equation:

$$\begin{aligned} & \frac{1}{n} \cdot \left( \sum_{i:\bar{x}_i \leq z_k} (\bar{x}_i)^2 + \sum_{i:z_{k+1} \leq \underline{x}_i} \underline{x}_i^2 + n_k \cdot \alpha^2 \right) - \\ & \frac{1}{n^2} \cdot \left( \sum_{i:\bar{x}_i \leq z_k} \bar{x}_i + \sum_{i:z_{k+1} \leq \underline{x}_i} \underline{x}_i + n_k \cdot \alpha \right)^2 = V_0. \end{aligned} \quad (2.4)$$

Then:

- if none of the two roots of the above quadratic equation belongs to the zone, this zone is dismissed;
- if one or more roots belong to the zone, then for each of these roots, based on this  $\alpha$ , we compute the value

$$E_k(\alpha) = \frac{1}{n} \cdot \left( \sum_{i:\bar{x}_i \leq z_k} \bar{x}_i + \sum_{i:z_{k+1} \leq \underline{x}_i} \underline{x}_i + n_k \cdot \alpha \right). \quad (2.5)$$

- After that:

- if  $V_0 < V^-$ , we return the smallest of the values  $E_k(\alpha)$  as  $\underline{E}$ :  $\underline{E} = \min_{k,\alpha} E_k(\alpha)$ ;
- if  $V_0 < V^+$ , we return the largest of the values  $E_k(\alpha)$  as  $\bar{E}$ :  $\bar{E} = \max_{k,\alpha} E_k(\alpha)$ .

*Comment.* The correctness of this algorithm is proven in the special Proof section.

## 2.4 Computation Time of This Algorithm

Sorting  $2n$  numbers requires time  $O(n \cdot \log(n))$ .

Once the values are sorted, we can then go zone-by-zone, and perform the corresponding computations. A straightforward implementation of the above algorithm would require time  $O(n^2)$ : for each of  $2n$  zones, we need linear time to compute several sums of  $n$  numbers.

However, in reality, only the sum for the first zone requires linear time. Once we have the sums for each zone, computing the sum for the next zone requires changing a few terms – values  $x_j$  which changed status. Each value  $x_j$  changes once, so overall, to compute all these sums, we still need linear time.

Thus, after sorting, the algorithm requires only linear computations time  $O(n)$ . So, if the endpoints are already given to us as sorted, we only take linear time.

If we still need to sort, then we need time

$$O(n \cdot \log(n)) + O(n) = O(n \cdot \log(n)).$$

## 2.5 Simple Example

Let us illustrate the above algorithm on a simple example in which we have two intervals  $\mathbf{x}_1 = [-1, 0]$  and  $\mathbf{x}_2 = [0, 1]$ , and the bound  $V_0$  is equal to 0.16.

In this case, according to the above algorithm, we compute the values

$$E^- = \frac{1}{2} \cdot (-1 + 0) = -0.5;$$

$$V^- = \frac{1}{2} \cdot (((-1) - (-0.5))^2 + (0 - (-0.5))^2) = 0.25;$$

$$E^+ = \frac{1}{2} \cdot (0 + 1) = 0.5;$$

$$V^+ = \frac{1}{2} \cdot ((0 - 0.5)^2 + (1 - 0.5)^2) = 0.25.$$

Here,  $V_0 < V^-$  and  $V_0 < V^+$ , so for computing both bounds  $\underline{E}$  and  $\overline{E}$ , we need to consider different zones.

By sorting the 4 endpoints  $-1, 0, 0, 1$ , we get  $z_1 = -1 \leq z_2 = 0 \leq z_3 = 0 \leq z_4 = 1$ . Thus, here, we have 3 zones  $[z_1, z_2] = [-1, 0]$ ,  $[z_2, z_3] = [0, 0]$ , and  $[z_3, z_4] = [0, 1]$ .

1) For the first zone  $[z_1, z_2] = [-1, 0]$ , according to the above algorithm, we select  $x_2 = 0$  and  $x_1 = \alpha$ . To determine the value  $\alpha$ , we form the quadratic equation (2.4):

$$\frac{1}{2} \cdot (0^2 + \alpha^2) - \frac{1}{4} \cdot (0 + \alpha)^2 = V_0 = 0.16.$$

This equation is equivalent to

$$\frac{1}{2} \cdot \alpha^2 - \frac{1}{4} \cdot \alpha^2 = \frac{1}{4} \cdot \alpha^2 = 0.16,$$

hence  $\alpha^2 = 0.64$  and  $\alpha = \pm 0.8$ . Of the two roots  $\alpha = -0.8$  and  $\alpha = 0.8$ , only the first root belongs to the zone  $[-1, 0]$ . For this root, we compute the value (2.5):

$$E_1 = \frac{1}{2} \cdot (0 + \alpha) = \frac{1}{2} \cdot (0 + (-0.8)) = -0.4.$$

2) For the second zone  $[z_2, z_3] = [0, 0]$ , according to the above algorithm, we select  $x_1 = x_2 = 0$ . In this case, there is no need to compute  $\alpha$ , so we directly compute

$$E_2 = \frac{1}{2} \cdot (0 + 0) = 0.$$

3) For the third zone  $[z_3, z_4] = [0, 1]$ , according to the above algorithm, we select  $x_1 = 0$  and  $x_2 = \alpha$ . To determine the value  $\alpha$ , we form the quadratic equation (2.4):

$$\frac{1}{2} \cdot (0^2 + \alpha^2) - \frac{1}{4} \cdot (0 + \alpha)^2 = V_0 = 0.16.$$

This equation is equivalent to

$$\frac{1}{2} \cdot \alpha^2 - \frac{1}{4} \cdot \alpha^2 = \frac{1}{4} \cdot \alpha^2 = 0.16,$$

hence  $\alpha^2 = 0.64$  and  $\alpha = \pm 0.8$ . Of the two roots  $\alpha = -0.8$  and  $\alpha = 0.8$ , only the second root belongs to the zone  $[0, 1]$ . For this root, we compute the value (2.5):

$$E_3 = \frac{1}{2} \cdot (0 + \alpha) = \frac{1}{2} \cdot (0 + 0.8) = 0.4.$$

Here, we have a value  $E_k$  for all three zones, so we return

$$\underline{E} = \min(E_1, E_2, E_3) = -0.4;$$

$$\overline{E} = \max(E_1, E_2, E_3) = 0.4.$$

## 2.6 Proof of the Algorithm's Correctness

1°. Let us first show that it is sufficient to prove correctness for the case of the upper endpoint  $\overline{E}$ .

Indeed, one can easily see that if we replace the original values  $x_i$  with the new values  $x'_i = -x_i$ , then the mean changes sign  $E' = -E$  while the variance remains the same  $V' = V$ .

When each  $x_i$  is known with interval uncertainty  $x_i \in \mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ , the corresponding interval for  $x'_i = -x_i$  is equal to  $\mathbf{x}'_i = [-\overline{x}_i, -\underline{x}_i]$ . The resulting interval  $\mathbf{E}' = [\underline{E}', \overline{E}']$  for  $E'$  is similarly equal to  $[-\overline{E}, -\underline{E}]$ , so  $\overline{E}' = -\underline{E}$  and thus,  $\underline{E} = -\overline{E}'$ .

Thus, if we know how to compute the upper endpoint  $\overline{E}$  for an arbitrary set of intervals  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , we can compute  $\underline{E}$  or a given set of intervals  $\mathbf{x}_1 = [\underline{x}_1, \overline{x}_1], \dots, \mathbf{x}_n = [\underline{x}_n, \overline{x}_n]$  as follows:

- we compute  $n$  auxiliary intervals  $\mathbf{x}'_i = [-\overline{x}_i, -\underline{x}_i]$ ,  $i = 1, \dots, n$ ;
- we use the known algorithm to find the upper endpoint  $\overline{E}'$  for the range of the mean when  $x'_i \in \mathbf{x}'_i$  and  $V(x') \leq V_0$ ;
- we take  $\underline{E} = -\overline{E}'$ .

2°. Let us prove that the largest possible values  $\overline{E}$  is attained for some values  $x_i \in [\underline{x}_i, \overline{x}_i]$  for which  $V(x) \leq V_0$ .

Indeed, the variance function  $V(x_1, \dots, x_n)$  is continuous; thus, the set of all the values  $x = (x_1, \dots, x_n)$  for which  $V(x_1, \dots, x_n) \leq V_0$  is closed.

The box  $\mathbf{x}_1 \times \dots \times \mathbf{x}_n$  is closed and bounded and thus, compact. The set  $S$  of all the values  $x \in \mathbf{x}_1 \times \dots \times \mathbf{x}_n$  for which  $V(x) \leq V_0$  is a closed subset of a compact set and therefore, compact itself. A continuous function attains its maximum on a compact set at some point. In particular, this means that the function  $E(x)$  attains its maximum  $\bar{E}$  at some point  $x$ , i.e., that there exist values  $x = (x_1, \dots, x_n)$  for which  $E(x_1, \dots, x_n) = \bar{E}$ .

In the following text, we will consider these optimizing values.

3°. Let us prove that for the optimizing vector  $x$ , for all  $i$  for which we have  $x_i < E$ , we have  $x_i = \bar{x}_i$ .

Indeed, since  $V = M - E^2$ , where  $M \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^n x_i^2$ , we conclude that

$$\frac{\partial V}{\partial x_i} = \frac{\partial M}{\partial x_i} - \frac{\partial E^2}{\partial x_i} = \frac{\partial M}{\partial x_i} - 2 \cdot E \cdot \frac{\partial E}{\partial x_i}.$$

Here,  $\frac{\partial E}{\partial x_i} = \frac{1}{n}$ ,  $\frac{\partial M}{\partial x_i} = \frac{2x_i}{n}$ , and therefore,

$$\frac{\partial V}{\partial x_i} = \frac{2 \cdot (x_i - E)}{n}. \quad (2.6)$$

If we change only one value  $x_i$ , by replacing it with  $x_i + \Delta x_i$ , with a small  $\Delta x_i$ , the value of  $V$  changes by

$$\Delta V = \frac{\partial V}{\partial x_i} \cdot \Delta x_i + o(\Delta x_i) = \frac{2}{n} \cdot (x_i - E) \cdot \Delta x_i + o(\Delta x_i). \quad (2.7)$$

When  $x_i < E$ , i.e., when  $x_i - E < 0$ , then for small  $\Delta x_i > 0$ , we have a negative  $\Delta V$ , i.e., the variance decreases, while the mean  $E$  increases by  $\frac{1}{n} \cdot \Delta x_i > 0$ . Thus, if we had  $x_i < E$  and  $x_i \neq \bar{x}_i$  for some  $i$ , then we could, by slightly increasing  $x_i$ , further increase  $E$  while decreasing  $V$  (and thus, keeping the constraint  $V \leq V_0$ ). So, in this case, the vector  $x$  cannot be the one that maximizes  $E$  under the constraint  $V \leq V_0$ .

This conclusion proves that for the optimizing vector, when  $x_i < E$ , we have  $x_i = \bar{x}_i$ .

4°. Let us assume that an optimizing vector has a component  $x_i$  which is strictly inside the corresponding interval  $[\underline{x}_i, \bar{x}_i]$ , i.e., for which  $\underline{x}_i < x_i < \bar{x}_i$ . Due to Part 3 of this proof, we cannot have  $x_i < E$ , so we must have  $x_i \geq E$ . Let us prove that in this case,

- for every  $j$  for which  $E \leq x_j < x_i$ , we have  $x_j = \bar{x}_j$ , and
- for every  $k$  for which  $x_k > x_i$ , we have  $x_k = \underline{x}_k$ .

4.1°. Let us first prove that if  $x_i \in (\underline{x}_i, \bar{x}_i)$ , and  $E \leq x_j < x_i$ , then  $x_j = \bar{x}_j$ .

We will prove this by contradiction. Indeed, let us assume that we have  $E \leq x_j < x_i$  and  $x_j < \bar{x}_j$ . In this case, we can, in principle, slightly increase  $x_j$ , to  $x_j + \Delta x_j$  and slightly decrease  $x_i$ , to  $x_i - \Delta x_i$ , and still stay within the corresponding intervals  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . We select  $\Delta x_j$  and  $\Delta x_i$  in such a way that the resulting change  $\Delta V$  in the variance  $V$  is non-negative. Here,

$$\Delta V = \frac{\partial V}{\partial x_j} \cdot \Delta x_j - \frac{\partial V}{\partial x_i} \cdot \Delta x_i + o(\Delta x_i) + o(\Delta x_j). \quad (2.8)$$

Substituting the formula (2.6) for the derivative  $\frac{\partial V}{\partial x_j}$  into this formula, we conclude that

$$\begin{aligned} \Delta V = \frac{2}{n} \cdot ((x_j - E)\Delta x_j - (x_i - E) \cdot \Delta x_i) + \\ o(\Delta x_i) + o(\Delta x_j). \end{aligned} \quad (2.9)$$

Thus, for every  $\Delta x_j$ , to get  $\Delta V = 0$ , we select

$$\Delta x_i = \frac{x_j - E}{x_i - E} \cdot \Delta x_j + o(\Delta x_j). \quad (2.10)$$

For this selection, the variance does not change, but the mean  $E$  is changed by

$$\begin{aligned} \Delta E = \frac{1}{n} \cdot (\Delta x_j - \Delta x_i) = \left(1 - \frac{x_j - E}{x_i - E}\right) \cdot \Delta x_j + o(\Delta x_j) = \\ \frac{x_i - x_j}{x_i - E} \cdot \Delta x_j + o(\Delta x_j). \end{aligned} \quad (2.11)$$

Since  $x_j < x_i$ , for small  $\Delta x_j$ , we have  $\Delta E > 0$ . Thus, we can further increase the mean without violating the constraint  $V \leq V_0$ . This contradicts our assumption that  $x$  is the optimizing vector. Thus, when  $E < x_j < x_i$ , we cannot have  $x_j < \bar{x}_j$  – so we must have  $x_j = \bar{x}_j$ .

4.2°. Let us first prove that if  $x_i \in (\underline{x}_i, \bar{x}_i)$ ,  $E \leq x_i$ , and  $x_k > x_i$ , then  $x_k = \underline{x}_k$ .

Similarly, let us assume that we have  $x_k > x_i$  and  $x_k > \underline{x}_k$ . In this case, we can, in principle, slightly increase  $x_i$ , to  $x_i + \Delta x_i$  and slightly decrease  $x_k$ , to  $x_k - \Delta x_k$ , and still stay within the corresponding intervals  $\mathbf{x}_i$  and  $\mathbf{x}_k$ . We select  $\Delta x_i$  and  $\Delta x_k$  in such a way that the resulting change  $\Delta V$  in the variance  $V$  is non-negative. Here,

$$\begin{aligned} \Delta V &= \frac{\partial V}{\partial x_i} \cdot \Delta x_i - \frac{\partial V}{\partial x_k} \cdot \Delta x_k + o(\Delta x_i) + o(\Delta x_k) = \\ &= \frac{2}{n} \cdot ((x_i - E)\Delta x_i - (x_k - E) \cdot \Delta x_k) + o(\Delta x_i) + o(\Delta x_k). \end{aligned} \quad (2.12)$$

Thus, for every  $\Delta x_i$ , to get  $\Delta V = 0$ , we select

$$\Delta x_k = \frac{x_i - E}{x_k - E} \cdot \Delta x_i + o(\Delta x_i). \quad (2.13)$$

For this selection, the variance does not change, but the mean  $E$  is changed by

$$\begin{aligned} \Delta E &= \frac{1}{n} \cdot (\Delta x_i - \Delta x_k) = \left(1 - \frac{x_i - E}{x_k - E}\right) \cdot \Delta x_i + o(\Delta x_i) = \\ &= \frac{x_k - x_i}{x_k - E} \cdot \Delta x_i + o(\Delta x_i). \end{aligned} \quad (2.14)$$

Since  $x_k > x_i$ , for small  $\Delta x_i$ , we have  $\Delta E > 0$ . Thus, we can further increase the mean without violating the constraint  $V \leq V_0$ . This contradicts our assumption that  $x$  is the optimizing vector. Thus, when  $x_i < x_k$ , we cannot have  $x_k > \underline{x}_k$  – so we must have  $x_k = \underline{x}_k$ .

5°. Let us now consider the case when for all the components  $x_i \geq E$  of the optimizing vector  $x$ , we have either  $x_i = \underline{x}_i$  or  $x_i = \bar{x}_i$ . Let us show that in this case, all the values  $x_i$  for which  $x_i = \bar{x}_i$  are smaller than or equal to all the values  $x_j$  for which  $x_j = \underline{x}_j$ .

We will prove this statement by contradiction. Let us assume that there exists  $i$  and  $j$  for which  $E \leq x_j < x_i$ ,  $x_j = \underline{x}_j$  and  $x_i = \bar{x}_i$ . In this case, we can slightly increase the value  $x_j$ , to  $x_j + \Delta x_j$ , and slightly decrease the value  $x_i$ , to  $x_i - \Delta x_i$ , and still stay within the corresponding intervals. Similarly to Part 4 of this proof, for every  $\Delta x_j > 0$ , to get  $\Delta V = 0$ , we must select

$$\Delta x_i = \frac{x_j - E}{x_i - E} \cdot \Delta x_j + o(\Delta x_j). \quad (2.15)$$

For this selection, the variance does not change, but the mean  $E$  is changed by

$$\begin{aligned} \Delta E &= \frac{1}{n} \cdot (\Delta x_j - \Delta x_i) = \left(1 - \frac{x_j - E}{x_i - E}\right) \cdot \Delta x_j + o(\Delta x_j) = \\ &= \frac{x_i - x_j}{x_i - E} \cdot \Delta x_j + o(\Delta x_j). \end{aligned} \quad (2.16)$$

Since  $x_j < x_i$ , for small  $\Delta x_j$ , we have  $\Delta E > 0$ . Thus, we can further increase the mean without violating the constraint  $V \leq V_0$ . This contradicts our assumption that  $x$  is the optimizing vector. So, when  $E \leq x_j < x_i$ , we cannot have  $x_j = \underline{x}_j$  and  $x_i = \bar{x}_i$ .

This contradiction proves that all the values  $x_i$  for which  $x_i = \bar{x}_i$  are indeed smaller than or equal to all the values  $x_j$  for which  $x_j = \underline{x}_j$ .

6°. Due to Parts 3, 4, and 5 of this proof, there exists a threshold value  $\alpha$  such that

- for all  $j$  for which  $x_j < \alpha$ , we have  $x_j = \bar{x}_j$ , and
- for all  $k$  for which  $x_k > \alpha$ , we have  $x_k = \underline{x}_k$ .

Indeed, in the case described in Part 4, as such  $\alpha$ , we can take the value  $x_i$  that is strictly inside the corresponding interval  $\mathbf{x}_i$ . In the case described in Part 5, since all the upper endpoints from the optimizing vector are smaller than or equal to all the lower endpoints, we can take any value  $\alpha$  between the largest of the optimal values  $\bar{x}_j$  and smallest of the optimal values  $\underline{x}_j$ .

7°. Let us show that because of the property proven in Part 6, once we know to which zone  $\alpha$  belongs, we can uniquely determine all the components  $x_j$  of the corresponding vector  $x$  – a candidate for the optimal vector.

7.1°. Indeed, if  $\bar{x}_j < \alpha$ , then, since we have  $x_j < \bar{x}_j$ , we get  $x_j < \alpha$ . Thus, due to Part 6, we have  $x_j = \bar{x}_j$ .

7.2°. If  $\alpha < \underline{x}_j$ , then, since we have  $\underline{x}_j < x_j$ , we get  $\alpha < x_j$ . Thus, due to Part 6, we have  $x_j = \underline{x}_j$ .

7.3°. Let us now consider the remaining case when neither of the above two conditions is satisfied and thus, we have  $\underline{x}_j \leq \alpha \leq \bar{x}_j$ .

In this case, we cannot have  $x_j < \alpha$ , because then, due to Part 6, we would have  $x_j = \bar{x}_j$  and thus,  $\bar{x}_j < \alpha$ , which contradicts the inequality  $\alpha \leq \bar{x}_j$ .

Similarly, we cannot have  $\alpha < x_j$ , because then, due to Part 6, we would have  $x_j = \underline{x}_j$  and thus,  $\alpha < \underline{x}_j$ , which contradicts the inequality  $\underline{x}_j \leq \alpha$ .

Thus, the only possible value here is  $x_j = \alpha$ .

7.3°. Overall, we conclude that for each  $\alpha$ , we get exactly the arrangement formulated in our algorithm.

8°. Let us prove that when  $V_0 < V^+$ , then the maximum is attained when  $V = V_0$ .

Let us prove this by contradiction. Let us assume that  $V_0 < V^+$  and that the maximum of  $E$  is attained for some vector  $x = (x_1, \dots, x_n)$ , with  $x_i \in [\underline{x}_i, \bar{x}_i]$ , for which  $V(x) < V_0$ .

Since  $V < V_0 < V^+$ , we have  $V(x) < V^+ = V(\bar{x}_1, \dots, \bar{x}_n)$ . Thus,  $x = (x_1, \dots, x_n) \neq \bar{x} \stackrel{\text{def}}{=} (\bar{x}_1, \dots, \bar{x}_n)$  – otherwise, we would get  $V(x) = V(\bar{x}) = V^+$ . So, there exists an index  $i$  for which  $x_i \neq \bar{x}_i$ . Since  $x_i \in [\underline{x}_i, \bar{x}_i]$ , this means that  $x_i < \bar{x}_i$ . Thus, we can increase  $x_i$  by a small positive value  $\varepsilon > 0$ , to a new value  $x'_i = x_i + \varepsilon > x_i$ , and still remain inside the interval  $[\underline{x}_i, \bar{x}_i]$ .

The function  $V(x_1, \dots, x_n)$  describing covariance continually depends on  $x_i$ . Since  $V(x) < V_0$ , for sufficiently small  $\varepsilon$ , we will have  $V(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n) < V_0$ . Thus, the new vector still satisfies the constraint – but for this new vector, the mean is larger (by  $\varepsilon/n > 0$ ) than for the original vector  $x$ .

This contradicts our assumption that the mean  $E(x)$  of the vector  $x$  is the largest possible under the given constraint  $V \leq V_0$ .

The above contradiction shows that when  $V_0 < V^+$ , then for the optimizing vector  $x$ , we have  $V(x) = V_0$ . This fact enables us to determine  $\alpha$  – as do in the algorithm – by solving the equation  $V(x(\alpha)) = V_0$ , where  $x(\alpha)$  is a vector corresponding to the given  $\alpha$ .

Correctness is proven.

# Chapter 3

## Estimating Covariance for the Privacy Case under Interval Uncertainty

In this chapter we show how to have a feasible algorithm for estimating covariance for privacy case under interval uncertainty. The main results of this chapter were published in [5].

### 3.1 Formulation of the Problem

As we have mentioned in Chapter 2, in general, the problems of computing the range of variance and covariance under interval uncertainty are, in general, NP-hard. It is therefore desirable to find practically important cases when we can compute these ranges in feasible (i.e., polynomial) time.

It turns out (see, e.g., [10, 11, 13]) that the range of variance can be computed in polynomial time when the intervals come from the need to preserve privacy in statistical databases, i.e., when we set some threshold values  $t_0, \dots, t_N$  and assume that each interval  $[\underline{x}_i, \bar{x}_i]$  coincides with one of the intervals  $[t_j, t_{j+1}]$ .

In this chapter, we show that for privacy case, the range of covariance can also be computed in polynomial time, when we set up  $x$ -thresholds  $t_j^{(x)}$  and  $y$ -thresholds  $t_j^{(y)}$  and assume that:

- each  $x$ -interval  $[\underline{x}_i, \bar{x}_i]$  coincides with one of the intervals  $[t_j^{(x)}, t_{j+1}^{(x)}]$ , and

- each  $y$ -interval  $[\underline{y}_i, \bar{y}_i]$  coincides with one of the intervals  $[t_j^{(y)}, t_{j+1}^{(y)}]$ .

## 3.2 Analysis of the Problem

**Reducing maximum to minimum.** When we change the sign of  $y_i$ , the covariance changes sign as well:  $C_{xy}(x_i, -y_i) = -C_{xy}(x_i, y_i)$ . Thus, for the ranges, we get

$$\mathbf{C}_{xy}(\mathbf{x}_i, -\mathbf{y}_i) = -\mathbf{C}_{xy}(\mathbf{x}_i, \mathbf{y}_i).$$

Since the function  $z \rightarrow -z$  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  $\bar{z}$ , the range of  $-z$  is  $[-\bar{z}, -\underline{z}]$ . Therefore,  $\underline{C}_{xy}(x_i, -y_i) = -\bar{C}_{xy}(x_i, y_i)$ .

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

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

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

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

If it attains its minimum at the point  $x = \underline{x}$ , then we cannot have  $\frac{df}{dx} < 0$ , because then, for some point  $x + \Delta x \in [\underline{x}, \bar{x}]$ , we would have a smaller value of  $f(x)$ . Thus, in this case, we must have  $\frac{df}{dx} \geq 0$ .

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

For the maximum, a similar thing happens. If  $f(x)$  attains its 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 must have  $\frac{df}{dx} \leq 0$ . Finally, if a function  $f(x)$  attains its maximum at the point  $x = \bar{x}$ , then we must have  $\frac{df}{dx} \geq 0$ .

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

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

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

- if  $x_i = \underline{x}_i$ , then  $y_i \geq E_y$ ;
- if  $x_i = \bar{x}_i$ , then  $y_i \leq E_y$ ;
- if  $\underline{x}_i < x_i < \bar{x}_i$ , then  $y_i = E_y$ .

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

- we cannot have  $x_i = \underline{x}_i$  — because then we would have  $y_i \geq E_y$ ; and
- we cannot have  $\underline{x}_i < x_i < \bar{x}_i$  — because then, we would have  $y_i = E_y$ .

So, if  $\bar{y}_i < E_y$ , the only remaining option for  $x_i$  is  $x_i = \bar{x}_i$ .

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

- we cannot have  $x_i = \bar{x}_i$  — because then we would have  $y_i \leq E_y$ ; and
- we cannot have  $\underline{x}_i < x_i < \bar{x}_i$  — because then, we would have  $y_i = E_y$ .

So, if  $E_y < \underline{y}_i$ , the only remaining option for  $x_i$  is  $x_i = \underline{x}_i$ .

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

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

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

So, if:

- the interval  $\mathbf{x}_i$  is either completely to the left or to the right of  $E_x$ , and
- the interval  $\mathbf{y}_i$  is either completely to the left or to the right of  $E_y$ ,

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

For example, if we know:

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

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

What if one of the intervals, e.g.,  $\mathbf{x}_i$ , is fully to the left or fully to the right of  $E_x$ , but  $\mathbf{y}_i$  contains  $E_y$  inside? For example, if  $\bar{x}_i < E_x$ , this means that  $y_i = \bar{y}_i$ . Since  $E_y$  is inside the interval  $[\underline{y}_i, \bar{y}_i]$ , this means that  $\underline{y}_i \leq E_y \leq \bar{y}_i$  and thus,  $E_y \leq y_i$ . If  $E_y < y_i$ , then, as we have shown earlier, we get  $x_i = \underline{x}_i$ . One can show that the same conclusion holds when  $y_i = E_y$ . So, in this case, we also have a single pair  $(x_i, y_i)$  where the minimum can be attained:  $x_i = \underline{x}_i$  and  $y_i = \bar{y}_i$ .

The only remaining case is when:

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

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

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

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

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

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

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

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

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

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

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

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

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

Similarly, the new values of  $E_x$  and  $E_y$  are:

$$E_x'' = \frac{1}{n} \cdot \sum_{j=1} x_j'' = \frac{1}{n} \cdot \sum_{j=1} x_j' + \frac{1}{n} \cdot \Delta = \frac{\Delta}{n};$$

$$E_y'' = \frac{1}{n} \cdot \sum_{j=1} y_j'' = \frac{1}{n} \cdot \sum_{j=1} y_j' - \frac{1}{n} \cdot \Delta = -\frac{\Delta}{n}.$$

Thus,

$$E_x'' \cdot E_y'' = \frac{\Delta}{n} \cdot \left(-\frac{\Delta}{n}\right) = -\frac{\Delta^2}{n^2},$$

and so,

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

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

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

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

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

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

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

- this interval may coincide with the corresponding  $x$ -range; in this case,  $E_x \in \mathbf{x}_i$ ;
- this interval may be completely to the left of this range; in this case,  $\bar{x}_i \leq E_x$ ; and

- this interval may be completely to the right of this range; in this case,  $E_x \leq \underline{x}_i$ .

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

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

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

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

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

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

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

So, we arrive at the following algorithm.

### 3.3 Resulting Algorithm

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

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

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

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

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

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

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

$$[\underline{C}_{xy}, \overline{C}_{xy}] = \{C(x_1, \dots, x_n, y_1, \dots, y_n) :$$

$$x_1 \in \mathbf{x}_1, \dots, x_n \in \mathbf{x}_n, y_1 \in \mathbf{y}_1, \dots, y_n \in \mathbf{y}_n\},$$

where

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

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

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

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

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

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

- the interval  $\mathbf{x}_i$  coincides with the  $x$ -range; we will denote this option by  $X^0$ ;

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Proof of correctness.** We know that for the minimizing vector  $(x_1, \dots, x_n, y_1, \dots, y_n)$ , the pair  $(E_x, E_y)$  must be contained in one of the  $N_x \cdot N_y$  cells.

We have already shown that for each cell, if the pair  $(E_x, E_y)$  is contained in this cell, then the corresponding minimizing values  $x_i$  and  $y_i$  – at which the covariance  $C_{xy}$  attains its smallest value  $\underline{C}_{xy}$  – will be as above. Thus, the actual minimizing value will be analyzed when we analyze the corresponding cell.

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

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

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

and  $N_y$   $z$ -ranges

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

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

### 3.4 Computation Time of the Algorithm

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

Each new computation differs from the previous one by a single change in  $\sum x_i \cdot y_i$  and a single change in estimating  $E_x \sim \sum x_i$  and  $E_y \sim \sum y_i$ . Thus, each new computation requires a constant time  $O(1)$ , and so, for each cell, the total computation time is  $O(n)$ . Thus, for all  $N_x \cdot N_y$  cells, we need time

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

**Discussion.** Usually, the number of  $x$ -ranges and the number of  $y$ -ranges are fixed. In this case, what we have is a *linear-time* algorithm.

Clearly, it is not possible to compute covariance faster than in linear time: we need to take into account all  $n$  data points, and processing each data point requires at least one computation.

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

# Chapter 4

## Estimating Correlation under Interval Uncertainty

### 4.1 Introduction

As we have mentioned in Chapter 1, in general, computing the range of correlation under interval uncertainty is NP-hard [2, 3, 13, 14, 17].

The problem of estimating correlation under interval uncertainty is formulated and analyzed in [16]; in that paper, this problem is formulated and solved as an optimization problem. For reasonably small  $n$ , the corresponding optimization algorithms work well [16]. However, since the problem is NP-hard, the computation time becomes infeasible when  $n$  is large.

In this chapter, we show that while we cannot have an efficient algorithm for computing both bounds  $\underline{\rho}$  and  $\bar{\rho}$ , we can effectively compute (at least) one of the bounds. Specifically, we show that we can compute  $\bar{\rho}$  when  $\bar{\rho} > 0$  and we can compute  $\underline{\rho}$  when  $\underline{\rho} < 0$ . This means that, in the case of a non-degenerate interval  $[\underline{\rho}, \bar{\rho}]$  (i.e.,  $\underline{\rho} < \bar{\rho}$ ):

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

## 4.2 Main Result and the Corresponding Algorithm

**Main result.** *There exists a polynomial-time algorithm that, given  $n$  pairs of intervals  $[\underline{x}_i, \bar{x}_i]$  and  $[\underline{y}_i, \bar{y}_i]$ , computes (at least) one of the endpoint of the interval  $[\underline{\rho}, \bar{\rho}]$  of possible values of the correlation  $\rho$ :*

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

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

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

Since the function  $z \rightarrow -z$  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  $\bar{z}$ , the range of  $-z$  is  $[-\bar{z}, -\underline{z}]$ . So, for the endpoints of the ranges, we get

$$\begin{aligned} \bar{\rho}([\underline{x}_1, \bar{x}_1], \dots, [\underline{x}_n, \bar{x}_n], -[\underline{y}_1, \bar{y}_1], \dots, -[\underline{y}_n, \bar{y}_n]) = \\ -\underline{\rho}([\underline{x}_1, \bar{x}_1], \dots, [\underline{x}_n, \bar{x}_n], [\underline{y}_1, \bar{y}_1], \dots, [\underline{y}_n, \bar{y}_n]), \end{aligned}$$

where

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

So, if we know how to compute the largest value  $\bar{\rho}$  when this value is positive, we can then compute the smallest value  $\underline{\rho}$  when this value is negative, as

$$\begin{aligned} \underline{\rho}([\underline{x}_1, \bar{x}_1], \dots, [\underline{x}_n, \bar{x}_n], [\underline{y}_1, \bar{y}_1], \dots, [\underline{y}_n, \bar{y}_n]) = \\ -\bar{\rho}([\underline{x}_1, \bar{x}_1], \dots, [\underline{x}_n, \bar{x}_n], [-\bar{y}_1, -\underline{y}_1], \dots, [-\bar{y}_n, -\underline{y}_n]). \end{aligned}$$

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

**Algorithm.** For each  $i$  from 1 to  $n$ , the corresponding box  $[\underline{x}_i, \bar{x}_i] \times [\underline{y}_i, \bar{y}_i]$  has four vertices:  $(\underline{x}_i, \underline{y}_i)$ ,  $(\underline{x}_i, \bar{y}_i)$ ,  $(\bar{x}_i, \underline{y}_i)$ , and  $(\bar{x}_i, \bar{y}_i)$ . 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  $\bar{\rho} = 1$ . 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  $[\underline{x}_i, \bar{x}_i] \times [\underline{y}_i, \bar{y}_i]$  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  $[\underline{x}_i, \bar{x}_i] \times [\underline{y}_i, \bar{y}_i]$ , 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  $x_i = \bar{x}_i$ . On the resulting segment  $\{\bar{x}_i\} \times [\underline{y}_i, \bar{y}_i]$ , 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  $y_i = \bar{y}_i$ ;
  - 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  $\{\underline{x}_i\} \times [\underline{y}_i, \bar{y}_i]$ , 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  $y_i = \bar{y}_i$ ;
  - 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  $[\underline{x}_i, \bar{x}_i] \times \{\underline{y}_i\}$ , 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  $x_i = \bar{x}_i$ ;
    - 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  $y_i = \bar{y}_i$ . On the resulting segment  $[\underline{x}_i, \bar{x}_i] \times \{\bar{y}_i\}$ , 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  $x_i = \bar{x}_i$ ;
    - 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  $\rho$  based on the corresponding values  $x_i$  and  $y_i$ .

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

*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  $\bar{x}_i$ ;
- 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  $K_x \stackrel{\text{def}}{=} \frac{1}{k_x} = \frac{V_x}{C}$ .

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 known value  $\bar{y}_i$ ;
- 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  $K_y \stackrel{\text{def}}{=} \frac{1}{k_y} = \frac{V_y}{C}$ .

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  $E_x = \frac{1}{n} \cdot \sum_{i=1}^n x_i$  is transformed into equating a linear combination of  $E_x$ ,  $K_x$ , and  $K_x \cdot E_y$ , to zero;
- the equation  $E_y = \frac{1}{n} \cdot \sum_{i=1}^n y_i$  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  $\leq 4$  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  $\leq 4$ . 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.

### 4.3 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  $\rho$  – by using the above explicit formula describing how the correlation  $\rho$  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  $[\underline{x}_i, \bar{x}_i] \times [\underline{y}_i, \bar{y}_i]$ , 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  $\rho \leq 1$ , this means that in this case,  $\bar{\rho} = 1$ .

**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  $[\underline{x}, \bar{x}]$  attains its maximum either at one of its endpoints, or in some internal point of the interval. If it attains its 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  $x + \Delta x \in [\underline{x}, \bar{x}]$ , we would have a larger value of  $f(x)$ . Thus, in this case, we must have  $\frac{df}{dx} \leq 0$ .

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

**Computing the corresponding derivatives.** We are interested in the values  $x_i$  and  $y_i$  for which the correlation  $\rho$  attains maximum. To use the above facts, let us find the partial derivatives of  $\rho$  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  $\sigma_x$  and  $\sigma_y$ . 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  $\rho$ .

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  $\rho = \frac{C}{\sigma_x \cdot \sigma_y}$ , since  $\sigma_y$  does not depend on  $x_i$ , we get

$$\begin{aligned} \frac{\partial \rho}{\partial x_i} &= \frac{1}{\sigma_y} \cdot \frac{\partial}{\partial x_i} \left( \frac{C}{\sigma_x} \right) = \frac{1}{\sigma_y} \cdot \frac{\frac{\partial C}{\partial x_i} \cdot \sigma_x - C \cdot \frac{\partial \sigma_x}{\partial x_i}}{\sigma_x^2} = \\ &= \frac{1}{\sigma_y \cdot \sigma_x^2 \cdot n} \cdot \left[ (y_i - E_y) \cdot \sigma_x - C \cdot \frac{x_i - E_x}{\sigma_x} \right]. \end{aligned}$$

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

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

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

**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  $\rho$  attains its largest possible value  $\bar{\rho} > 0$ . Then, according to the above facts from calculus, we have one of the three possible situations:

- $x_i \in (\underline{x}_i, \bar{x}_i)$  and  $\frac{\partial \rho}{\partial x_i} = 0$ , i.e.,  $y_i = E_y + k_x \cdot (x_i - E_x)$ ;
- $x_i = \underline{x}_i$  and  $\frac{\partial \rho}{\partial x_i} \leq 0$ , i.e.,  $y_i \leq E_y + k_x \cdot (x_i - E_x)$ ;
- $x_i = \bar{x}_i$  and  $\frac{\partial \rho}{\partial x_i} \geq 0$ , i.e.,  $y_i \geq E_y + k_x \cdot (x_i - E_x)$ .

Here,  $k_x$  has the same sign as the correlation, so  $k_x > 0$ . Let us now consider possible locations of the box  $[\underline{x}_i, \bar{x}_i] \times [\underline{y}_i, \bar{y}_i]$  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  $[\underline{x}_i, \bar{x}_i] \times [\underline{y}_i, \bar{y}_i]$  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  $y_i \in [\underline{y}_i, \bar{y}_i]$  and  $x_i \in [\underline{x}_i, \bar{x}_i]$ . In this case, we cannot have  $x_i \in (\underline{x}_i, \bar{x}_i)$  and  $x_i = \underline{x}_i$ , so we must have  $x_i = \bar{x}_i$ .

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

- $y_i \in (\underline{y}_i, \bar{y}_i)$  and  $\frac{\partial \rho}{\partial y_i} = 0$ , i.e.,  $x_i = E_x + k_y \cdot (y_i - E_y)$ ;
- $y_i = \underline{y}_i$  and  $\frac{\partial \rho}{\partial y_i} \leq 0$ , i.e.,  $x_i \leq E_x + k_y \cdot (y_i - E_y)$ ;
- $y_i = \bar{y}_i$  and  $\frac{\partial \rho}{\partial y_i} \geq 0$ , i.e.,  $x_i \geq E_x + k_y \cdot (y_i - E_y)$ .

Here,  $k_y$  has the same sign as the correlation, so  $k_y > 0$ . Let us now consider possible locations of the segment  $\{\bar{x}_i\} \times [\underline{y}_i, \bar{y}_i]$  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  $y_i \in [\underline{y}_i, \bar{y}_i]$ . In this case, we cannot have  $y_i \in (\underline{y}_i, \bar{y}_i)$  and we cannot have  $y_i = \bar{y}_i$ , 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  $y_i \in [\underline{y}_i, \bar{y}_i]$ . In this case, we cannot have  $y_i \in (\underline{y}_i, \bar{y}_i)$  and we cannot have  $y_i = \underline{y}_i$ , so we must have  $y_i = \bar{y}_i$ .

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  $y'_i \in [\underline{y}_i, \bar{y}_i]$ . 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  $\bar{y}_i$ .

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  $\underline{y}_i \neq y'_i$ , 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  $x_i \leq E_x + k_y \cdot (\underline{y}_i - E_y)$ , and therefore, the maximum cannot be attained for  $y_i = \underline{y}_i$ .

1.3.3°. If  $\bar{y}_i \neq y'_i$ , then  $y'_i < \bar{y}_i$ , and thus,

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

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

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  $[\underline{x}_i, \bar{x}_i] \times [\underline{y}_i, \bar{y}_i]$  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  $y_i \in [\underline{y}_i, \bar{y}_i]$  and  $x_i \in [\underline{x}_i, \bar{x}_i]$ . In this case, we cannot have  $x_i \in (\underline{x}_i, \bar{x}_i)$  and we cannot have  $x_i = \bar{x}_i$ , 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  $y_i = \bar{y}_i$  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,  $y - E_y = \frac{1}{k_y} \cdot (x - E_x)$ . 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  $\frac{1}{k_y}$  for the  $y$ -line. Thus, these lines coincide if and only if  $k_x = \frac{1}{k_y}$ , i.e., if and only if  $k_x \cdot k_y = 1$ .

In general,  $\rho \leq 1$ . Here,  $\rho = \frac{C}{\sigma_x \cdot \sigma_y} = \frac{C}{\sqrt{V_x} \cdot \sqrt{V_y}}$ ; thus,  $\rho = \sqrt{k_x \cdot k_y}$ , so  $k_x \cdot k_y \leq 1$ . If  $k_x \cdot k_y < 1$ , then  $k_x \cdot k_y \neq 1$  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 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.

# Chapter 5

## Concluding Remarks

In many applications, it is important to estimate the values of the statistical characteristics such as mean, variance, covariance, and correlation based on the results of observations and/or measurements. In these estimates, it is important to take into account measurement errors. Often, the only information that we have about each measurement error  $\Delta x$  is an upper bound  $\Delta$  on its value  $|\Delta x| \leq \Delta$ , we do not have any information about the probabilities of different values from the corresponding interval  $[-\Delta, \Delta]$ . In such cases, as a result of the measurement, the only information that we gain about the actual (unknown) value  $x$  of the corresponding quantity is that this value belongs to the interval

$$[x, \bar{x}] = [\tilde{x} - \Delta, \tilde{x} + \Delta].$$

It is therefore necessary to estimate the range of each statistical characteristic – mean, variance, covariance, and correlation – under the corresponding interval uncertainty.

It is known that the problem of estimating the range of the mean under interval uncertainty is easy to solve – but it becomes more complex if we take into account that there may be constraints that limit possible values  $x_1, \dots, x_n$  from the corresponding intervals. In contrast to the mean, the problems of computing the ranges of variance, covariance, and correlation under interval uncertainty are, in general, NP-hard. It is therefore desirable to find practically useful classes of problems for which feasible algorithms are possible. Such classes are known for the case of variance: e.g., it is known that one of the endpoints of the variance range is feasible to compute, and that both endpoints can be feasibly computed in the privacy case, when the intervals come not from measurement uncertainty, but from the need to preserve privacy in statistical databases.

In this thesis, we show that the range for mean can be feasibly computed even in the presence of variance-related constraints. We also provide feasible algorithms for computing the range of covariance in the privacy case, and for computing one of the endpoints of the range of correlation.

# 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*, 2002, Vol. 33, No. 2, pp. 108–118.
- [3] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, “Exact Bounds on Finite Populations of Interval Data,” *Reliable Computing*, 2005, Vol. 11, No. 3, pp. 207–233.
- [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] A. Jalal-Kamali, V. Kreinovich, and L. Longpré, “Estimating Covariance for Privacy Case under Interval (and Fuzzy)”, In: *Proceedings of the World Conference on Soft Computing*, San Francisco, CA, May 23–26, 2011.
- [6] A. Jalal-Kamali, L. Longpre, and M. Koshelev, “Estimating Mean under Interval Uncertainty and Variance Constraint”, In: *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society NAFIPS’2011*, El Paso, Texas, March 18–20, 2011.

- [7] 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.
- [8] M. Koshelev, A. Jalal-Kamali, L. Longpré, “Estimating sample mean under interval uncertainty and constraint on sample variance”, *International Journal of Approximate Reasoning*, 2011, Vol. 52, No. 8, pp. 1136–1146.
- [9] V. Kreinovich, “Reliability Analysis for Aerospace Applications: Reducing Over-Conservative Expert Estimates in the Presence of Limited Data,” In: S. O. Kuznetsov and D. 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.
- [10] 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.
- [11] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos, “Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity,” *Reliable Computing*, Vol. 12, No. 6, pp. 471–501, 2006.
- [12] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM Press, Philadelphia, Pennsylvania, 2009.
- [13] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, 2011.

- [14] R. Osegueda, V. Kreinovich, L. Potluri, and R. Al'o, "Non-destructive testing of aerospace structures: granularity and data mining approach," *Proc. FUZZ-IEEE'2002*, Honolulu, Hawaii, May 12–17, 2002, Vol. 1, pp. 685–689.
- [15] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer Verlag, New York, 2005.
- [16] 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.
- [17] G. Xiang, *Fast Algorithms for Computing Statistics under Interval Uncertainty, with Applications to Computer Science and to Electrical and Computer Engineering*, PhD Dissertation, Department of Computer Science, University of Texas at El Paso, 2007.
- [18] G. Xiang, M. Ceberio, and V. Kreinovich, "Computing Population Variance and Entropy under Interval Uncertainty: Linear-Time Algorithms," *Reliable Computing*, 2007, Vol. 13, No. 6, pp. 467–488.

# Curriculum Vitae

Ali Jalal-Kamali was born on February 28, 1986. The youngest child of Hossein Jalal-Kamali and Fatemeh Ali-Ahmadi, he graduated from Bu-Ali High School, Kerman, Iran, in the spring of 2002. He entered Shahid-Bahonar University of Kerman in the Spring of 2003, and received his bachelor's degree in Computer Science in the summer of 2007. While pursuing his bachelor's degree in Computer Science, he worked as a System Requirement Analyst, and as a Tutor / Instructor for English and Computer Basics. After graduating from University of Kerman, in 2008 he started work as the project coordinator in Huawei Co. Ltd. and later on in 2009 was promoted to project manager.

In the fall of 2009, he entered the Graduate School of The University of Texas at El Paso (UTEP). While pursuing a Master's degree in Computer Science he worked at UTEP as a Teaching Assistant, as a Research Assistant, and as an Instructor for CS1420 "Computer Programming for Scientists and Engineers" at the UTEP. He is a member of the Sigma Xi, the Scientific Research Society.

Permanent address: 200 Wallington Dr., Apt 206

El Paso, Texas 79902