Computing Covariance and Correlation in Optimally Privacy-Protected Statistical Databases: Feasible Algorithms

Joshua Day[1], Ali Jalal-Kamali[2], and Vladik Kreinovich[2]

[1]Department of Computer Science, University of Wisconsin at Whitewater,
Whitewater, WI 53190, USA, dayja10@uww.edu
[2]Department of Computer Science, University of Texas at El Paso,
El Paso, TX 79968, USA, ajalalkamali@miners.utep.edu, vladik@utep.edu

Abstract. In many real-life situations, e.g., in medicine, it is necessary to process data while preserving the patients' confidentiality. One of the most efficient methods of preserving privacy is to replace the exact values with intervals that contain these values. For example, instead of an exact age, a privacy-protected database only contains the information that the age is, e.g., between 10 and 20, or between 20 and 30, etc. Based on this data, it is important to compute correlation and covariance between different quantities. For privacy-protected data, different values from the intervals lead, in general, to different estimates for the desired statistical characteristic. Our objective is then to compute the range of possible values of these estimates.

Algorithms for effectively computing such ranges have been developed for situations when intervals come from the original surveys, e.g., when a person fills in whether his or her age is between 10 or 20, between 20 and 30, etc. These intervals, however, do not always lead to an optimal privacy protection; it turns out that more complex, computer-generated "intervalization" can lead to better privacy under the same accuracy, or, alternatively, to more accurate estimates of statistical characteristics under the same privacy constraints. In this paper, we extend the existing efficient algorithms for computing covariance and correlation based on privacy-protected data to this more general case of interval data.

1 Formulation of the Problem

Need for processing data in statistical databases. Often, we collect data for the purpose of finding possible dependencies between different quantities. For example, we collect all possible information about the medical patients with the hope of finding out which factors affect different illnesses and which factors affect the success of different cures. The resulting collection of records $r_i = (r_{i1}, \ldots, r_{ip})$, $1 \le i \le n$, is known as a *statistical database* since typically, statistical methods are used for look for possible dependencies; see, e.g., [7]. These statistical methods are usually based on computing statistical characteristics such as mean

$E_j = \frac{1}{n} \cdot \sum_{i=1}^{n} r_{ij}$, variance $V_j = \frac{1}{n} \cdot \sum_{i=1}^{n} (r_{ij} - E_j)^2$, standard deviation $\sigma_j = \sqrt{V_j}$, covariance $C_{jk} = \frac{1}{n} \cdot \sum_{i=1}^{n} (r_{ij} - E_j) \cdot (r_{ik} - E_k)$, and correlation $\rho_{jk} = \frac{C_{jk}}{\sigma_j \cdot \sigma_k}$.

Need for privacy protection. In many real-life situations, e.g., in medicine, it is necessary to process data while preserving the patients' confidentiality.

One of the most efficient methods of preserving privacy is to replace the exact values with intervals that contain these values. For example, instead of an exact age, a privacy-protected database only contains the information that the age is, e.g., between 10 and 20, or between 20 and 30, etc.

In general, for each of p variables x_i, $1 \le i \le p$, we fix some thresholds $t_{i,1} < t_{i,2} < \ldots < t_{i,n_i}$ (e.g., 0, 10, 20, 30, ..., for age), and replace each original value x_i with the range $[t_{i,k}, t_{i,k+1}]$ that contains this value. In the above example, the actual age of 19 will be replaced by the range $[10, 20]$.

Need to process corresponding interval data. Based on this interval data, it is important to compute the values of different statistical characteristics such as correlation and covariance between different quantities.

For privacy-protected data, for each statistical characteristic $C(v_1, \ldots, v_m)$, different values v_i from the given intervals $[\underline{v}_i, \overline{v}_i]$ lead, in general, to different estimates $C(v_1, \ldots, v_m)$. Thus, it is necessary to compute the range of possible values of these estimates:

$$C([\underline{v}_1, \overline{v}_1], \ldots, [\underline{v}_m, \overline{v}_m]) \stackrel{\text{def}}{=}$$

$$\{C(v_1, \ldots, v_m) : v_1 \in [\underline{v}_1, \overline{v}_1], \ldots, v_m \in [\underline{v}_m, \overline{v}_m]\}. \qquad (1)$$

What was known before. For most statistical characteristics, the problem of computing the range (1) under *general* interval uncertainty is NP-hard; see, e.g., [6]. However, for the above-described privacy-related case, feasible algorithms are possible for computing many statistical characteristics, in particular, covariance and correlation; see, e.g., [2–6].

Need to go beyond the threshold-based "intervalization". In the above threshold-based "intervalization", we replace each data point $r = (r_1, \ldots, r_p)$ with a box

$$b = [\underline{b}_1, \overline{b}_1] \times \ldots \times [\underline{b}_p, \overline{b}_p] \qquad (2)$$

formed by the corresponding threshold intervals $[\underline{b}_i, \overline{b}_i]$. The larger the boxes, the wider the resulting interval (1) – i.e., the less accurate our estimates of the corresponding statistical characteristics. From this viewpoint, the boxes b should be as narrow as possible. On the other hand, if they are too narrow, e.g., if some box contains only one record, then the privacy of this record is not well-protected. To properly protect privacy, we need to make sure for some sufficiently large integer K, each box b contains at least K records (this is called

K*-anonymity*; see, e.g., [8]), and that for each variable x_i, there are at least ℓ different values of this variable coming from records within this box (this is called ℓ*-diversity*); see, e.g., [1].

Boxes do not have to come from thresholds. The only reasonable restriction is that they should form a *subdivision* in the sense that no two boxes should have a common interior point. Under the privacy-motivated restrictions of K-anonymity and ℓ-diversity, we must look for a subdivision into boxes which leads to the narrowest possible range $C([\underline{v}_1, \overline{v}_1], \ldots, [\underline{v}_p, \overline{v}_p])$ of the desired characteristic. It turns out (see, e.g., [9, 10]) that to attain this narrowest range, we need to use a general subdivision into boxes which is more complex than the above threshold-based one. Namely, in the above threshold-based subdivision into boxes, if two records $(r_1, r_2, \ldots)$ and $(r'_1, r'_2, \ldots)$ have the same value of r_1 (i.e., if $r'_1 = r_1$), then the corresponding boxes have the same x_1-interval $[\underline{b}_1, \overline{b}_1]$. In other words, the selection of the x_1-interval of the corresponding box depends only on the value r_1 and does not depend on the values of all other quantities $r_2, \ldots$

In contrast, in the optimal subdivision into boxes, the same value of r_1, depending on the values of other quantities $r_2, \ldots$, we may need boxes with different x_1-intervals. For example, if for some $r_2, \ldots$, there are more records around the point $(r_1, r_2, \ldots)$, then, in the optimal subdivision into boxes, these records are assigned to a narrower box, with narrower x_1-intervals. On the other hand, for the same value r_1 and different values $r'_2, \ldots$, there may be much fewer records around the point $(r_1, r'_2, \ldots)$. In this case, in the optimal subdivision into boxes, these new records records are assigned to a wider box, with a wider x_1-interval.

Resulting problem and what we do in this paper. Since the optimal intervalization goes beyond a simple threshold-based one, it is necessary to extend algorithms for estimating covariance and correlation to such optimal intervalization. Such algorithms are presented in this paper.

2 Analysis of the Problem

First comment: computing the upper endpoint $\overline{C}_{jk}$ can be reduced to computing the lower endpoint $\underline{C}_{jk}$. One can easily check that if we replace each value r_{ik} with its opposite $r'_{ik} = -r_{ik}$, then the covariance C_{jk} changes sign: $C'_{jk} = -C_{jk}$. As a result, if we replace each original interval $[\underline{r}_{ik}, \overline{r}_{ik}]$ with its opposite $[-\overline{r}_{ik}, -\underline{r}_{ik}]$, then the resulting range is the opposite to the original range: $[\underline{C}'_{jk}, \overline{C}'_{jk}] = [-\overline{C}_{jk}, -\underline{C}_{jk}]$. This means, in particular, that $\underline{C}'_{jk} = -\overline{C}_{jk}$ and therefore, that $\overline{C}_{jk} = -\underline{C}'_{jk}$.

Thus, if we know how to compute lower endpoints, we can compute the lower endpoint $\underline{C}'_{jk}$ for the modified database, and then compute $\overline{C}_{jk}$ as $\overline{C}_{jk} = -\underline{C}'_{jk}$.

Because of this reduction, in the following text, we will only consider the problem of computing the lower endpoint $\underline{C}_{jk}$.

Known facts from calculus: reminder. Each statistical characteristic $C(v_1, \ldots, v_m)$ is a continuous function of its variables. It is known that the range of a continuous function on a connected box $[\underline{v}_1, \overline{v}_1] \times \ldots \times [\underline{v}_m, \overline{v}_m]$ is an interval $[\underline{C}, \overline{C}]$ whose endpoints are the smallest possible value $\underline{C}$ of the function $C(v_1, \ldots, v_m)$ on the box and its largest value $\overline{C}$. It is also known that for each continuous function on a closed box, its minimum and its maximum are attained at some points.

When a function $C(v_1, \ldots, v_m)$ attains its minimum on the box at a point $(v_1^{\min}, \ldots, v_i^{\min}, \ldots, v_m^{\min})$, this means, in particular, that for every i, the one-variable function $f(v_i) \stackrel{\text{def}}{=} C(v_1^{\min}, \ldots, v_{i-1}^{\min}, v_i, v_{i+1}^{\min}, \ldots, v_m^{\min})$ attains its minimum on the interval $[\underline{v}_i, \overline{v}_i]$ at $v_i = v_i^{\min}$.

In general, a function $f(x)$ of one variable attains its minimum on an interval $[\underline{x}, \overline{x}]$ either inside this interval or at one of its endpoints $\underline{x}$ or $\overline{x}$. If the function $f(x)$ attains its minimum at an inside point, then its derivative at this point is known to be equal to 0: $f'(x^{\min}) = 0$. If $f(x)$ attains its minimum at $\underline{x}$, then we should have $f'(\underline{x}) \geq 0$ because otherwise, if we had $f'(\underline{x}) < 0$, then, for a small Δx, we would have $f(\widetilde{x} + \Delta x) < f(\underline{x})$, which contradicts to our assumption that the value $f(\underline{x})$ is the smallest. Similarly, if the function $f(x)$ attains its minimum at $\overline{x}$, we should have $f'(\overline{x}) \leq 0$.

Let us apply these facts to minimizing covariance. For covariance, as one can easily check, $\dfrac{\partial C_{jk}}{\partial r_{ij}} = \dfrac{1}{n} \cdot (r_{ik} - E_k)$ and $\dfrac{\partial C_{jk}}{\partial r_{ik}} = \dfrac{1}{n} \cdot (r_{ij} - E_j)$. Thus, for the values $r_{ij}^{\min}$ and $r_{ik}^{\min}$ at which the minimum of C_{jk} is attained, we have one of the three options:

- either $\underline{r}_{ij} < r_{ij}^{\min} < \overline{r}_{ij}$ and $\dfrac{\partial C_{jk}}{\partial r_{ij}} = 0$, i.e., $r_{ik}^{\min} = E_k$;
- or $r_{ij}^{\min} = \underline{r}_{ij}$ and $r_{ik}^{\min} \geq E_k$;
- or $r_{ij}^{\min} = \overline{r}_{ij}$ and $r_{ik}^{\min} \leq E_k$.

Thus:

- if $r_{ik}^{\min} > E_k$, then the first and third cases are impossible, so we must have $r_{ij}^{\min} = \underline{r}_{ij}$;
- if $r_{ik}^{\min} < E_k$, then the first and second cases are impossible, so we must have $r_{ij}^{\min} = \overline{r}_{ij}$.

Therefore, if $E_k < \underline{r}_{ik}$, then, due to $\underline{r}_{ik} \leq r_{ik}^{\min}$, we get $E_k < r_{ik}^{\min}$ and therefore, $r_{ij}^{\min} = \underline{r}_{ij}$. Similarly, if $\overline{r}_{ik} < E_k$, then $r_{ij}^{\min} = \overline{r}_{ij}$.

Likewise, if $r_{ij}^{\min} > E_j$, then $r_{ik}^{\min} = \underline{r}_{ik}$, and if $r_{ij}^{\min} < E_j$, then $r_{ik}^{\min} = \overline{r}_{ik}$. So, if $E_j < \underline{r}_{ij}$, then $r_{ik}^{\min} = \underline{r}_{ik}$, and if $\overline{r}_{ij} < E_j$, then $r_{ik}^{\min} = \overline{r}_{ik}$.

Thus, if we know the location of E_j in comparison to the interval $[\underline{r}_{ij}, \overline{r}_{ij}]$ and we know the location of E_k in comparison with the interval $[\underline{r}_{ik}, \overline{r}_{ik}]$, then, with one exception, we can uniquely determine the minimizing values $r_{ij}^{\min}$ and $r_{ik}^{\min}$. For example, if $E_k < \underline{r}_{ik}$ and $E_j < \underline{r}_{ij}$, then $r_{ij}^{\min} = \overline{r}_{ij}$ and $r_{ik}^{\min} = \overline{r}_{ik}$. If $E_k < \underline{r}_{ik}$ and $\underline{r}_{ij} \leq E_j \leq \overline{r}_{ij}$, then $r_{ij}^{\min} = \overline{r}_{ij} \geq E_j$, hence $r_{ik}^{\min} = \overline{r}_{ik}$.

The only exception is when $E_j \in [\underline{r}_{ij}, \overline{r}_{ij}]$ and $E_k \in [\underline{r}_{ik}, \overline{r}_{ik}]$. In this case, minimizing over r_{ij}, we have three calculus-motivated options:

- the first option is $r_{ik}^{\min} = E_k$;
- the second option is $r_{ij}^{\min} = \underline{r}_{ij}$ and $r_{ik}^{\min} \ge E_k$;
- the third option is $r_{ij}^{\min} = \overline{r}_{ij}$ and $r_{ik}^{\min} \le E_k$.

These conditions describe a set of possible pairs $(r_{ij}^{\min}, r_{ik}^{\min})$, a set formed by three line segments.

Similarly, minimizing over r_{ik}, we have three other calculus-motivated options:

- the first option is $r_{ij}^{\min} = E_j$;
- the second option is $r_{ik}^{\min} = \underline{r}_{ik}$ and $r_{ij}^{\min} \ge E_j$;
- the third option is $r_{ik}^{\min} = \overline{r}_{ik}$ and $r_{ij}^{\min} \le E_j$,

which define a new three-segment set. The actual pair $(r_{ij}^{\min}, r_{ik}^{\min})$ belongs to both these sets and thus, belongs to their intersection. This intersection consists of three points: $(\underline{r}_{ij}, \overline{r}_{ik})$, $(\overline{r}_{ij}, \underline{r}_{ik})$, and (E_j, E_k).

Let us show that the minimum cannot be attained at a point (E_j, E_k). Indeed, let us show that if for some small $\Delta \ne 0$, we replace the value $r_{ij} = E_j$ with a new value $r'_{ij} = E_j + \Delta$ and the value $r_{ik} = E_k$ with a new value $r'_{ik} = E_k - \Delta$, then the covariance will decrease – which shows that the minimum is not attained when $r_{ij} = E_j$ and $r_{ik} = E_k$. To show this, we will use a known equivalent expression for the covariance $C_{jk} = M - E_j \cdot E_k$, where $M \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} r_{ij} \cdot r_{ik}$. When we replace the values r_{ij} and r_{ik} with the new values r'_{ij} and r'_{ik}, then the mean E_j is replaced with $E'_j = E_j + \frac{\Delta}{n}$, the mean E_k is replaced with $E'_k = E_k - \frac{\Delta}{n}$. The product $r_{ij} \cdot r_{ik} = E_j \cdot E_k$ is replaced with

$$(E_j + \Delta)(E_k - \Delta) = E_j \cdot E_k - \Delta \cdot E_j + \Delta \cdot E_k - \Delta^2.$$

Thus, the quantity M is replaced with $M' = M - \frac{1}{n} \cdot \Delta \cdot E_j + \frac{1}{n} \cdot \Delta \cdot E_k - \frac{1}{n} \cdot \Delta^2$. Hence, the new expression for the covariance takes the form

$$C'_{jk} = M' - E'_j \cdot E'_k = M - \frac{1}{n} \cdot \Delta \cdot E_j + \frac{1}{n} \cdot \Delta \cdot E_k - \frac{1}{n} \cdot \Delta^2 - \left(E_j + \frac{\Delta}{n}\right) \cdot \left(E_k + \frac{\Delta}{n}\right).$$

After opening parentheses, we can see that the terms proportional to $\Delta \cdot E_j$ and $\Delta \cdot E_k$ cancel out, so we get $C'_{jk} = C_{jk} - \frac{1}{n} \cdot \Delta^2 + \frac{1}{n^2} \cdot \Delta^2 = C_{jk} - \frac{n-1}{n^2} \cdot \Delta^2 < C_{jk}$. This proves that when the box b contains the point (E_j, E_k), then we have only two options for the minimizing values of r_{ij} and r_{ik}.

Towards an algorithm. In the privacy-protected database, boxes form a subdivision, so for each possible location of the pair (E_j, E_k), there is at most one box that contains this pair. This box contains several records; let us denote their number by n_b. In the minimizing selection, some of the pairs $(r_{ij}^{\min}, r_{ik}^{\min})$ are equal to $(\underline{r}_{ij}, \overline{r}_{ik})$ and some are equal to $(\overline{r}_{ij}, \underline{r}_{ik})$. Covariance does not change if we re-order the records; thus, when computing covariance, we only care about how many of n_b records are equal to $(\underline{r}_{ij}, \overline{r}_{ik})$; let us denote this number by m_b. One can easily check that M, E_j, and E_k are linear functions of m_b; thus, the covariance $C_{jk} = M - E_j \cdot E_k$ is a quadratic function of m_b: $C_{jk} = C_2 \cdot m_b^2 + C_1 \cdot m_b + C_0$, for known values C_i.

To find the smallest possible value of C_{jk}, we want to find a value $m_b = 0, 1, \ldots, n_b$ for which this expression is the smallest possible. This can be done by using the known properties of a quadratic function $C_2 \cdot m_b^2 + C_1 \cdot m_b + C_0$:

- when $C_2 > 0$, it decreases when $m_b \le -\dfrac{C_1}{2C_2}$ and increases after that;
- when $C_2 < 0$, it increases when $m_b \le -\dfrac{C_1}{2C_2}$ and decreases after that;
- when $C_2 = 0$, it increases if $C_1 > 0$ and decreases if $C_1 < 0$.

On the interval where this expression is increasing, we take the smallest possible value of m_b; on the interval where this expression is decreasing, we take the largest possible value of m_b.

Towards an algorithm: final touch. What is important is where the values E_j and E_k are in comparison with the endpoints of the corresponding intervals $[\underline{r}_{ij}, \overline{r}_{ij}]$ and $[\underline{r}_{ik}, \overline{r}_{ik}]$. Thus, to find possible ranges of E_j, we can sort all the endpoints $\underline{r}_{ij}$ and $\overline{r}_{ij}$ of the x_j-intervals of different boxes into an increasing sequence $T_{j,1} < T_{j,2} < \ldots$, and consider all possible "small boxes" $b = [T_{j,i_j}, T_{j,i_j+1}] \times [T_{k,i_k}, T_{j,i_k+1}]$. Thus, we arrive at the following algorithm for computing the lower endpoint $\underline{C}_{jk}$ of the range of covariance.

3 Algorithm for Computing Covariance

What is given. We are given a finite collection of B boxes $b_a = [\underline{b}_{a1}, \overline{b}_{a1}] \times \ldots \times [\underline{b}_{ap}, \overline{b}_{ap}]$, $1 \le a \le B$. These boxes form a subdivision, i.e., no two boxes have a common interior point. For each of these boxes, we are given the number n_a of records corresponding to this box. We are also given the indices j and k for which we want to find the range of covariance values.

Algorithm. First, we sort all $2B$ j-endpoints $\underline{b}_{aj}$ and $\overline{b}_{aj}$ of all B boxes into an increasing sequence $T_{j,1} < T_{j,2} < \ldots$, and form $\le 2B$ "small" j-intervals $[T_{j,i_j}, T_{j,i_j+1}]$.

Then, we similarly sort all $2B$ k-endpoints $\underline{b}_{ak}$ and $\overline{b}_{ak}$ of all B boxes into an increasing sequence $T_{k,1} < T_{k,2} < \ldots$, and form $\le 2B$ "small" k-intervals $[T_{k,i_k}, T_{k,i_k+1}]$. After that, we form "small boxes" by considering all possible pairs $b = [T_{j,i_j}, T_{j,i_j+1}] \times [T_{k,i_k}, T_{j,i_k+1}]$ of a small j-interval and a small k-interval. In our algorithms, we will analyze these small boxes one by one.

Let us now consider computations corresponding to a fixed small box b. As we have shown, once the small box $b = [\underline{b}_j, \overline{b}_j] \times [\underline{b}_k, \overline{b}_k]$ is fixed, then for almost all original boxes (except for the original box b_{a_0} that contains b), we can uniquely determine the minimizing values $r_{ij}^{\min}$ and $r_{ik}^{\min}$:

- if $\overline{b}_j \le \underline{b}_{aj}$ and $\overline{b}_k \le \underline{b}_{ak}$, then $r_{ij}^{\min} = \underline{b}_{aj}$ and $r_{ik}^{\min} = \underline{b}_{ak}$;
- if $\overline{b}_j \le \underline{b}_{aj}$ and $\underline{b}_{ak} \le \underline{b}_k \le \overline{b}_k \le \overline{b}_{ak}$, then $r_{ij}^{\min} = \overline{b}_{aj}$ and $r_{ik}^{\min} = \underline{b}_{ak}$;
- if $\overline{b}_j \le \underline{b}_{aj}$ and $\overline{b}_{ak} \le \underline{b}_k$, then $r_{ij}^{\min} = \overline{b}_{aj}$ and $r_{ik}^{\min} = \underline{b}_{ak}$;
- if $\overline{b}_{aj} \le \underline{b}_j$ and $\overline{b}_k \le \underline{b}_{ak}$, then $r_{ij}^{\min} = \underline{b}_{aj}$ and $r_{ik}^{\min} = \overline{b}_{ak}$;
- if $\overline{b}_{aj} \le \underline{b}_j$ and $\underline{b}_{aj} \le \underline{b}_k \le \overline{b}_k \le \overline{b}_{ak}$, then $r_{ij}^{\min} = \underline{b}_{aj}$ and $r_{ik}^{\min} = \overline{b}_{ak}$;
- if $\overline{b}_{aj} \le \underline{b}_j$ and $\overline{b}_{ak} \le \underline{b}_k$, then $r_{ij}^{\min} = \overline{b}_{aj}$ and $r_{ik}^{\min} = \overline{b}_{ak}$;
- if $\underline{b}_{aj} \le \underline{b}_j \le \overline{b}_j \le \overline{b}_{aj}$ and $\overline{b}_k \le \underline{b}_{ak}$, then $r_{ij}^{\min} = \underline{b}_{aj}$ and $r_{ik}^{\min} = \overline{b}_{ak}$;
- if $\underline{b}_{aj} \le \underline{b}_j \le \overline{b}_j \le \overline{b}_{aj}$ and $\overline{b}_{ak} \le \underline{b}_k$, then $r_{ij}^{\min} = \overline{b}_{aj}$ and $r_{ik}^{\min} = \underline{b}_{ak}$.

This way, for each of the boxes b_a $(a \ne a_0)$, we can compute this box's contributions to the expressions M, E_j, and E_k as, correspondingly,

$$\frac{n_a}{n} \cdot r_{ij}^{\min} \cdot r_{ik}^{\min}, \quad \frac{n_a}{n} \cdot r_{ij}^{\min}, \text{ and } \frac{n_a}{n} \cdot r_{ik}^{\min}.$$

For the box $b_{a_0} = [\underline{b}_{a_0 j}, \overline{b}_{a_0 j}] \times [\underline{b}_{a_0 k}, \overline{b}_{a_0 k}]$, the corresponding contributions take the form

$$\frac{m_{a_0}}{n} \cdot \underline{b}_{a_0 j} \cdot \overline{b}_{a_0 k} + \frac{n_{a_0} - m_{a_0}}{n} \cdot \overline{b}_{a_0 j} \cdot \underline{b}_{a_0 k},$$

$$\frac{m_{a_0}}{n} \cdot \underline{b}_{a_0 j} + \frac{n_{a_0} - m_{a_0}}{n} \cdot \overline{b}_{a_0 j}, \text{ and } \frac{m_{a_0}}{n} \cdot \overline{b}_{a_0 k} + \frac{n_{a_0} - m_{a_0}}{n} \cdot \underline{b}_{a_0 k},$$

with an unknown m_{a_0}. By adding the contributions corresponding to different boxes and forming $C_{jk} = M - E_j \cdot E_k$, we get an expression for C_{jk} which is quadratic in m_{a_0}. By using techniques described in the previous section, we can compute the minimum of this expression over all possible integer values m_{a_0} from 0 to n_{a_0}. This minimum $C_{jk}(b)$ is the smallest possible value of the covariance under the assumption that the pair (E_j, E_k) belongs to the small box b.

To find the desired value $\underline{C}_{jk}$, we can then compute the smallest of the values $C_{jk}(b)$ corresponding to all possible small boxes b.

Computational time for this algorithm. Sorting takes time $O(B \cdot \log(B))$. After sorting, we get $\le 2B$ j-intervals and $\le 2B$ k-intervals, so we get $O(B^2)$ small boxes – pairs of such intervals.

In the main part of the algorithm, for each of $O(B^2)$ small boxes b and for each of B original boxes b_a, we need finitely many computational steps. Thus, the total number of computational steps for the main part is bounded by $O(B^2) \cdot B \cdot \text{const} = O(B^3)$. The total computation time is thus equal to $O(B \cdot \log(B)) + O(B^3)$, i.e., to $O(B^3)$. This algorithm requires cubic time and is, therefore, feasible.

Comment. According to [9], in some cases, better estimates for covariance come from weighted estimates $C^w_{jk} = \sum\limits_{i=1}^{n} w_i \cdot (r_{ij} - E^w_j) \cdot (r_{ik} - E^w_k)$, where

$$E^w_j = \sum_{i=1}^{n} w_i \cdot r_{ij}, \quad E^w_k = \sum_{i=1}^{n} w_i \cdot r_{ik},$$

and w_i are appropriate weights for which $w_i \geq 0$ and $\sum\limits_{i=1}^{n} w_i = 1$. The weight w_i of a record depends only on the box b_a that contains this record. In other words, for some values W_a, $w_i = W_a$ for all the records r_i from the box b_a. In these terms, the equality $\sum\limits_{i=1}^{n} w_i = 1$ means that $\sum\limits_{a} n_a \cdot W_a = 1$. The formula for C^w_{jk} can be represented in an equivalent form, as $C^w_{jk} = M^w - E^w_j \cdot E^w_k$, where $M^w_{jk} = \sum\limits_{i=1}^{n} w_i \cdot r_{ij} \cdot r_{ik}$.

An analysis similar to the one from Section 2 shows that, in effect, the algorithm from Section 3 can be applied for computing the range of this characteristic as well; the only difference is that after selecting the values $r^{\min}_{ij}$ and $r^{\min}_{ik}$, we need to use the weighted expressions M^w, E^w_j, and E^w_k instead of original equal-weight expressions for M, E_j, and E_k.

4 Algorithms for Computing Correlation

Correlation: reminder. The Pearson's correlation coefficient ρ describes the degree of dependence between the inputs: if the coefficient ρ is close to 1 or to -1, this means that there is a strong dependence; if this coefficient is close to 0, this means that most probably, there is no dependence.

Correlation under interval uncertainty: practical meaning of lower and upper bounds. Under interval uncertainty, instead of a single value ρ, we get an interval $\left[\underline{\rho}, \overline{\rho}\right]$ of possible values. For positive values ρ, the upper endpoint $\overline{\rho}$ describes to what extent it is *possible* that there is a dependence between the inputs, while the lower endpoint $\underline{\rho}$ describes to what extent, based on the available data, we can *guarantee* that there is a dependence. Similarly, for negative values ρ, the lower endpoint $\underline{\rho}$ describes to what extent it is *possible* that there is a dependence between the inputs, while the upper endpoint $\overline{\rho}$ describes to what extent, based on the available data, we can *guarantee* that there is a dependence.

Which endpoints are most important for statistical databases. As we have mentioned, one of the main purposes of statistical databases is to discover possible new dependencies – dependencies which can then be checked and utilized. From this viewpoint, the most important endpoints are: the upper endpoint for the positive correlation, and the lower endpoint for the negative correlation.

Computing correlation: what is known. The relative importance of different bounds is good news: while in general, computing correlation under interval uncertainty is NP-hard (see, e.g., [6]), a feasible (i.e., polynomial-time) algorithm is possible for computing the upper endpoint $\overline{\rho}$ for positive correlations and the lower endpoint $\underline{\rho}$ for negative correlations; see, e.g., [2].

The known algorithm is rather slow. This algorithm is polynomial-time: for inputs consisting of n records, its computation time is bounded by $O(n^5)$.

However, from the practical viewpoint, even for a small database with $n = 1000$ records, this means 10^{15} arithmetic operations: two weeks on a Gigaflop machine; for $n = 10^4$ records, this already means an unrealistic amount of 10^{20} operations.

For statistical databases with privacy-motivated boxes, the known algorithm can be made somewhat faster. In the algorithm from [2], we consider possible quadruples (pairs of pairs) of vertices. In the privacy-motivated case, we have $\leq 4B$ vertices, where B is the number of different boxes. Thus, the total number of quadruples of vertices is $O(B^4)$.

According to [2], once the quadruple is fixed, then, within each box b_a, we select the same optimizing values $r_{ij}^{\max}$ and $r_{ik}^{\max}$ (or $r_{ij}^{\min}$ and $r_{ik}^{\min}$) for all the records from this box. Thus, once the quadruple is fixed, we need to perform only finitely many computations within each box – and then, as we did for covariance, multiply the results by n_a. For each of $O(B^4)$ quadruples, we therefore need $O(B)$ computational steps, to the total of $O(B^4) \cdot O(B) = O(B^5)$.

This number of steps is still large, but since the number of boxes is much smaller than the number of records, this number of steps is much smaller than $O(n^5)$ – and thus, more realistic.

Acknowledgments

This work was supported by The National Center for Border Security and Immigration (NCBSI) at the University of Texas at El Paso, a Department of Homeland Security (DHS) Center of Excellence, by Grants 1 T36 GM078000-01 and 1R43TR000173-01 from the National Institutes of Health, by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by a grant N62909-12-1-7039 from the Office of Naval Research.

References

1. Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N.: A Framework for Efficient Data Anonymization under Privacy and Accuracy Constraints, ACM Transactions on Database Systems, 34(2), Article 9 (2009)
2. Jalal-Kamali, A., Kreinovich, V.: Estimating Correlation under Interval Uncertainty, Mechanical Systems and Signal Processing, 37, 43–53 (2013)

3. Jalal-Kamali, A., Kreinovich, V., Longpré, L.: Estimating Covariance for Privacy Case under Interval (and Fuzzy) Uncertainty, In: Yager, R.R., Reformat, M., Shahbazova, S., Ovchinnikov, S. (eds.), Proceedings of the World Conference on Soft Computing, San Francisco, CA, May 23-26, 2011 (2011)
4. Kreinovich, V., Longpré, L., Starks, S.A., Xiang, G., Beck, J., Kandathi, R., Nayak, A., Ferson, S., Hajagos, J.: Interval Versions of Statistical Techniques, with Applications to Environmental Analysis, Bioinformatics, and Privacy in Statistical Databases, Journal of Computational and Applied Mathematics, 199(2), 418–423 (2007)
5. Kreinovich, V., Xiang, G., Starks, S.A., Longpré, L., Ceberio, M., Araiza, R., Beck, J., Kandathi, R., Nayak, A., Torres, R., Hajagos, J.: Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity, Reliable Computing, 12(6), 471–501 (2006)
6. Nguyen, H.T., Kreinovich, V., Wu, B., Xiang, G.: Computing Statistics under Interval and Fuzzy Uncertainty. Springer Verlag, Berlin, Heidelberg (2012)
7. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, Boca Raton, Florida (2011)
8. Sweeney, L.: k-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-Based System, 10(5), 557–570 (2002)
9. Xiang, G., Ferson, S., Ginzburg, L., Longpré, L., Mayorga, E., Kosheleva, O.: Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics: Fuzzy-Motivated Approach, In: Proceedings of the Joint World Congress of the International Fuzzy Systems Association and Annual Conference of the North American Fuzzy Information Processing Society IFSA/NAFIPS'2013, Edmonton, Canada, June 24–28, 2013, pp. 611–616 (2013)
10. Xiang, G., Kreinovich, V.: Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics, In: Proceedings of the IEEE Symposium on Computational Intelligence for Engineering Solutions CIES'2013, Singapore, April 16–19, 2013, pp. 163–170 (2013)