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

1. 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 medical information about the patients to find out:
  - which factors affect different illnesses;
  - 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*.
- Statistical methods are used to look for possible dependencies.
- Most such methods use mean, variance, covariance, and correlation.

2. Need for Privacy Protection

- In many real-life situations, e.g., in medicine:
  - it is necessary to process data
  - while preserving the patients' confidentiality.
- *Idea:* replace the exact values with intervals that contain these values.
- For example, only check whether 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 thresholds $t_{i,1} < t_{i,2} < \ldots < t_{i,n_i}$ (e.g., 0, 10, 20, 30, $\ldots$, for age), and
  - replace each original value x_i with the range $[t_{i,k}, t_{i,k+1}]$ that contains this value.
- For example, age of 19 is replaced by $[10, 20]$.

3. Need to Process Corresponding Interval Data

- Different values v_i from the intervals 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:

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

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

- In general, the problem of computing this range is NP-hard.
- However, for the privacy case, feasible algorithms are possible, e.g., for covariance and correlation.

4. 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

$$b = [\underline{b}_1, \overline{b}_1] \times \ldots \times [\underline{b}_p, \overline{b}_p].$$

- The narrower the box, the more accurate our estimates of the corresponding statistical characteristics.
- But if boxes are too narrow, privacy is not protected.
- We need to guarantee that for some K, each box b contains at least K records (this is called K-*anonymity*).
- Optimal division-into-boxes under this constraint does not come from thresholds.
- For example, records with the same b_1 may end up in boxes with different intervals $[\underline{b}_1, \overline{b}_1]$.

5. Computing Upper Endpoints $\overline{C}_{jk}$ for Covariance Reduced to Computing Lower Endpoints

- If we replace each value r_{ik} with its opposite $r'_{ik} = -r_{ik}$, then the covariance C_{jk} changes sign: $C'_{jk} = -C_{jk}$.
- So, if we replace each original interval $[\underline{r}_{ik}, \overline{r}_{ik}]$ with its opposite $[-\overline{r}_{ik}, -\underline{r}_{ik}]$, then the range changes to

$$[\underline{C}'_{jk}, \overline{C}'_{jk}] = [-\overline{C}_{jk}, -\underline{C}_{jk}].$$

- Hence $\underline{C}'_{jk} = -\overline{C}_{jk}$ and $\overline{C}_{jk} = -\underline{C}'_{jk}$.
- Thus, if we know how to compute lower endpoints, we:
  - 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, we will only consider the problem of computing the lower endpoint $\underline{C}_{jk}$.

6. Known Facts from Calculus: Reminder

- 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 $f(x)$ attains its minimum at $\underline{x}$, then we should have $f'(\underline{x}) \geq 0$; otherwise:
  - if we had $f'(\underline{x}) < 0$,
  - then, for a small Δx, we have $f(\underline{x} + \Delta x) < f(\underline{x})$,
  - but $f(\underline{x})$ is the smallest value.
- Similarly, if the function $f(x)$ attains its minimum at $\overline{x}$, we have $f'(\overline{x}) \leq 0$.
- If min is attained inside, we should have $f'(x_{\rm min}) = 0$.

7. Applying These Facts to Covariance

- Covariance is $C_{jk} = \frac{1}{n} \cdot \sum_{i=1}^{n} (r_{ij} - E_j) \cdot (r_{ik} - E_k)$, so $\frac{\partial C_{jk}}{\partial r_{ij}} = \frac{1}{n} \cdot (r_{ik} - E_k)$ and $\frac{\partial C_{jk}}{\partial r_{ik}} = \frac{1}{n} \cdot (r_{ij} - E_j)$.
- Thus, for the minimizing values $r_{ij}^{\min}$ and $r_{ik}^{\min}$, we have:
  - either $\underline{r}_{ij} < r_{ij}^{\min} < \overline{r}_{ij}$ and $\frac{\partial C_{jk}}{\partial r_{ij}} = 0$, 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 $r_{ik}^{\min} \leq E_k$.
- This enables us, once we know where E_j and E_k are, to find where min is attained.

8. Resulting Algorithm

- *Given:* a finite collection of B disjoint boxes $b_a = [\underline{b}_{a1}, \overline{b}_{a1}] \times \ldots \times [\underline{b}_{ap}, \overline{b}_{ap}]$, $1 \le a \le B$.
- For each b_a, we know the number n_a of records $r \in b_a$.
- We want to compute C_{ij} for given j and k.
- First, we sort all $2B$ j-endpoints $\underline{b}_{aj}$ and $\overline{b}_{aj}$ of all B boxes into an increasing sequence $T_{j,1} < T_{j,2} < \ldots$
- We form $\le 2B$ "small" j-intervals $[T_{j,i_j}, T_{j,i_j+1}]$.
- Then, we sort all $2B$ k-endpoints $\underline{b}_{ak}$ and $\overline{b}_{ak}$ of all B boxes into an increasing sequence $T_{k,1} < T_{k,2} < \ldots$.
- We form $\le 2B$ "small" k-intervals $[T_{k,i_k}, T_{k,i_k+1}]$.
- 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 intervals.
- We analyze these small boxes one by one.

9. For Each Small Box b, If $(E_j, E_k) \in b$

- If $\overline{b}_j \le \underline{b}_{aj}$ & $\overline{b}_k \le \underline{b}_{ak}$, then $r^{\min}_{ij} = \underline{b}_{aj}$, $r^{\min}_{ik} = \underline{b}_{ak}$.
- If $\overline{b}_j \le \underline{b}_{aj}$ & $\underline{b}_{ak} \le \underline{b}_k \le \overline{b}_k \le \overline{b}_{ak}$, then $r^{\min}_{ij} = \overline{b}_{aj}$ and $r^{\min}_{ik} = \underline{b}_{ak}$.
- If $\overline{b}_j \le \underline{b}_{aj}$ & $\overline{b}_{ak} \le \underline{b}_k$, then $r^{\min}_{ij} = \overline{b}_{aj}$, $r^{\min}_{ik} = \underline{b}_{ak}$.
- If $\overline{b}_{aj} \le \underline{b}_j$ & $\overline{b}_k \le \underline{b}_{ak}$, then $r^{\min}_{ij} = \underline{b}_{aj}$, $r^{\min}_{ik} = \overline{b}_{ak}$.
- If $\overline{b}_{aj} \le \underline{b}_j$ & $\underline{b}_{aj} \le \underline{b}_k \le \overline{b}_k \le \overline{b}_{ak}$, then $r^{\min}_{ij} = \underline{b}_{aj}$ and $r^{\min}_{ik} = \overline{b}_{ak}$.
- If $\overline{b}_{aj} \le \underline{b}_j$ & $\overline{b}_{ak} \le \underline{b}_k$, then $r^{\min}_{ij} = \overline{b}_{aj}$, $r^{\min}_{ik} = \overline{b}_{ak}$.
- If $\underline{b}_{aj} \le \underline{b}_j \le \overline{b}_j \le \overline{b}_{aj}$ & $\overline{b}_k \le \underline{b}_{ak}$, then $r^{\min}_{ij} = \underline{b}_{aj}$ and $r^{\min}_{ik} = \overline{b}_{ak}$.
- If $\underline{b}_{aj} \le \underline{b}_j \le \overline{b}_j \le \overline{b}_{aj}$ & $\overline{b}_{ak} \le \underline{b}_k$, then $r^{\min}_{ij} = \overline{b}_{aj}$ and $r^{\min}_{ik} = \underline{b}_{ak}$.

10. Algorithm (cont-d)

- For the original b_{a_0} containing b, the minimizing record can be in one of the two opposite endpoints.
- This way, we get an expression for C_{jk} which is quadratic in the number of values m_{a_0} in one of the endpoints.
- We can easily find m_{a_0} that minimizes this expression.
- The minimum over all small boxes is the desired $\underline{C}_{jk}$.
- For each of $O(B^2)$ small boxes, we consider each of B original boxes, so this algorithm take time $O(B^3)$.
- A similar algorithm works for weighted estimate $C^w_{jk} = \sum\limits_{i=1}^n w_i \cdot (r_{ij} - E^w_j) \cdot (r_{ik} - E^w_k)$, where

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

11. Computing Correlation ρ

- The Pearson's correlation coefficient ρ describes the degree of dependence between the inputs:
  - if $\rho \approx 1$ or $\rho \approx -1$, there is a strong dependence.
  - if $\rho \approx 0$, there is no dependence.
- 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.
- For negative values ρ, the lower endpoint $\underline{\rho}$ describes to what extent it is *possible* that there is a dependence.
- One of the main purposes of statistical databases is to discover possible new dependencies.
- So, the most important endpoints are: $\overline{\rho}$ for $\rho > 0$ and $\underline{\rho}$ for $\rho < 0$.

12. Computing Correlation (cont-d)

- *Good news:* there is a feasible ($O(n^5)$) algorithm for computing $\overline{\rho}$ for $\rho > 0$ and $\underline{\rho}$ for $\rho < 0$.
- *Problem:* for $n = 10^4$ records, this means an unrealistic amount of 10^{20} operations.
- The known ρ-algorithm considers possible quadruples of vertices.
- In the privacy-motivated case, we have $\leq 4B$ vertices, so we have $O(B^4)$ quadruples.
- Once the quadruple is fixed, we need to perform only finitely many computations for each of B boxes.
- For each of $O(B^4)$ quadruples, we 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 much smaller than $O(n^5)$.

13. Acknowledgments

This work was supported

- by 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),
- by the National Science Foundation grant DUE-0926721, and
- by grant N62909-12-1-7039 from the Office of Naval Research.