Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics

Joshua Day

University of Wisconsin at Whitewater, REU at University of Texas at El Paso

Abstract

Often, when statistics are taken of data that needs to remain private (such as medical information), we use one of multiple methods of data anonymization to protect the confidentiality of an individual.

One common method is to replace exact values with intervals that contain these values, for example, that the recorded age of a person is between 10 and 20, 20 and 30, etc. The resulting thresholds, leading to boxes of possible values, can generally be used to calculate a range for particular statistics. This method still protects the privacy of individuals by not allowing accurate conclusions to be drawn about a particular characteristic.

With the intervals of data, when trying to compute a statistic, there will not be an exact answer, there is a range of possible values corresponding to different points from the boxes. Ranges resulting from thresholds fixed for each quantity are often to wide to be useful (e.g., [–1,1] for correlation). A known way to make these ranges narrower (and thus more useful) is to use non-overlapping boxes with their own particular intervals for the *x* and *y* characteristics.

We want to show how the existing algorithms for covariance and correlation can be extended to this general privacy-protected situation.

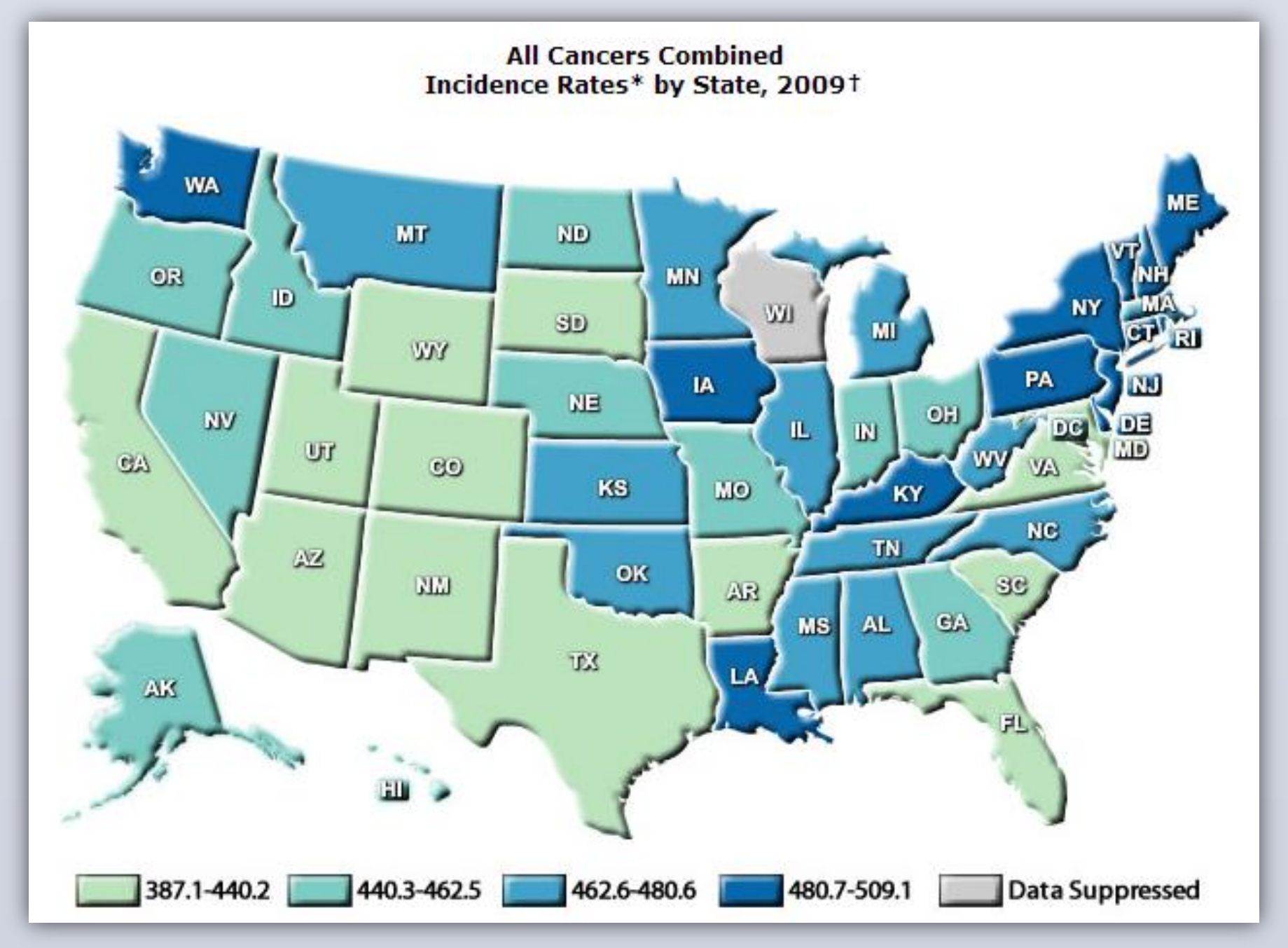
This results in efficient, polynomial-time algorithms capable of computing narrower ranges for covariance and correlation from privacy-motivated interval data.

Anonymizing Data

Protecting the privacy of a person (or a business) is important in statistical processing of medical information, census data, etc. Data can be anonymized by using encryption, perturbation, suppression, and/or generalization:

Methods	Examples
Data Encryption	"JohnDoe" → "@Gek1ds%#$"
Number and date variance	34 → 30
Nulling out specific fields or data sets	DATA → "--------"
Intervals	18 → [10,20]

In this research we use interval anonymization. Based on the interval data, it is necessary to compute the values of different statistical characteristics such as correlation and covariance between different quantities. For privacy-protected interval data, for each statistical characteristic $C(v_1, \ldots, v_m)$, different values v_i from the given 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:



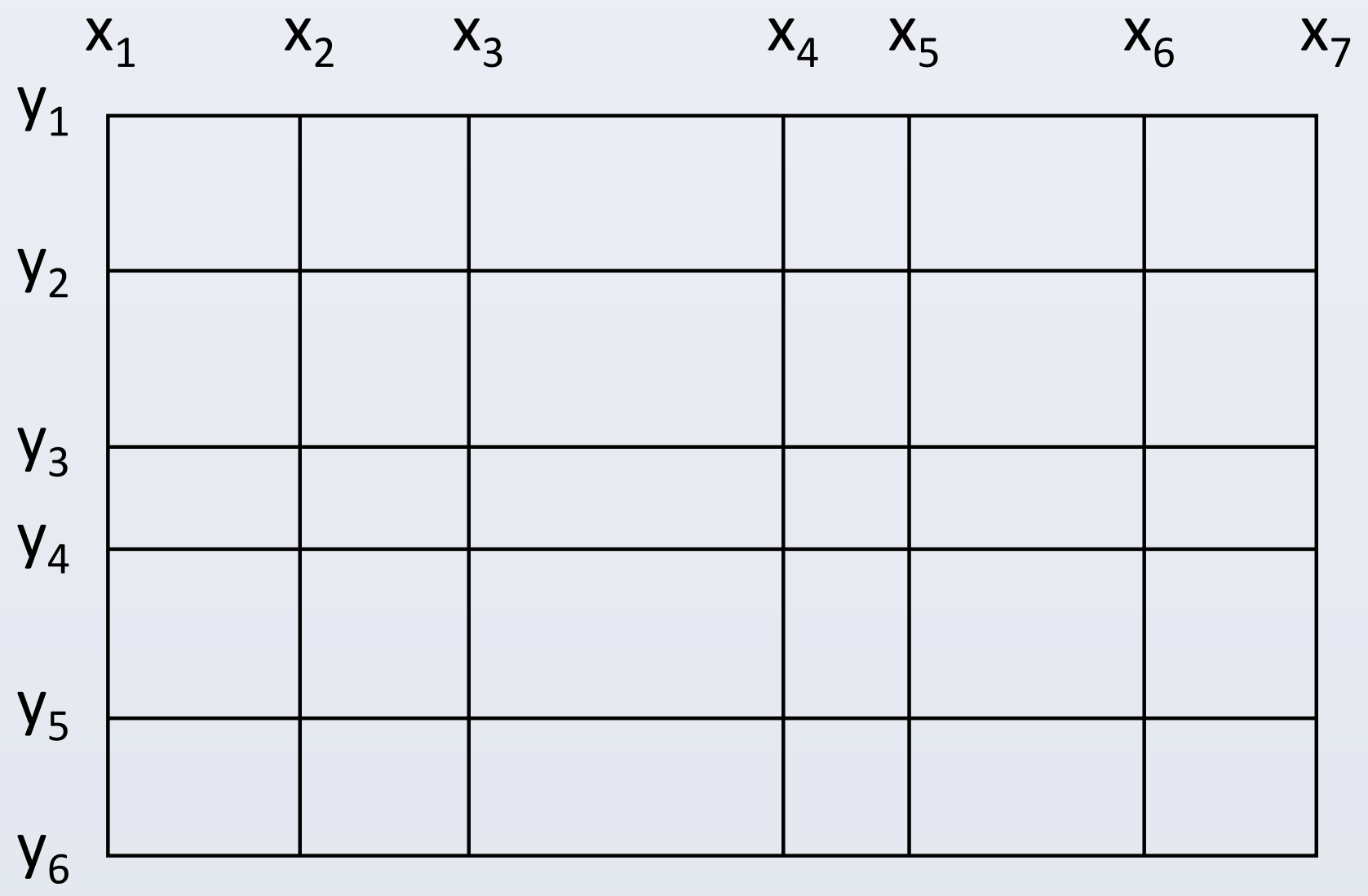
The above example of data anonymization showing statistics for all cancers combined by state is taken from the website of the Center for Disease Control (CDC).

Materials & Methods

Related Work

Much work has previously been done in the area of computations with intervals. Algorithms have been developed for computing a range for statistics such as covariance [1] and correlation [2] under interval uncertainty.

These algorithms are based on boxes formed by having thresholds for each x characteristic and y characteristic. The thresholds can neatly divide data boxes into rows and columns as shown below. However, this subdivision into boxes results in ranges which are too wide (e.g., [–1,1] for correlation).



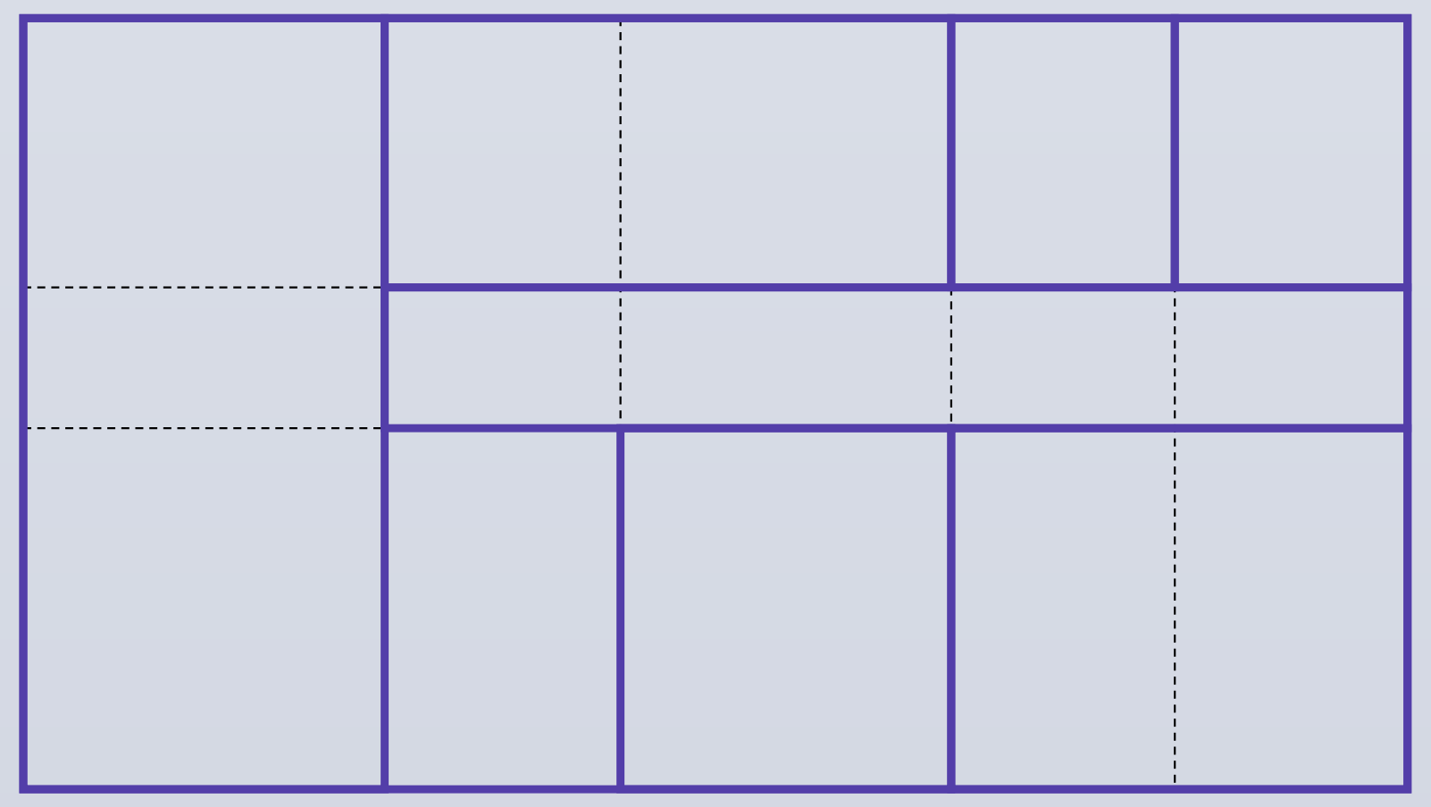
General Criteria for Anonymization

To properly protect privacy, we need to satisfy the following two criteria:

- **k-anonymity** each box B contains at least k records
- **ℓ-diversity** that for each variable x_i, there are at least ℓ different values of this variable coming from records within this box

Optimal Anonymization: What is Known

According to [3], different sized boxes lead to an optimal subdivision of data for anonymization in order to obtain the most accurate estimates for statistics. An example of what is meant by different sized boxes is shown outlined below.



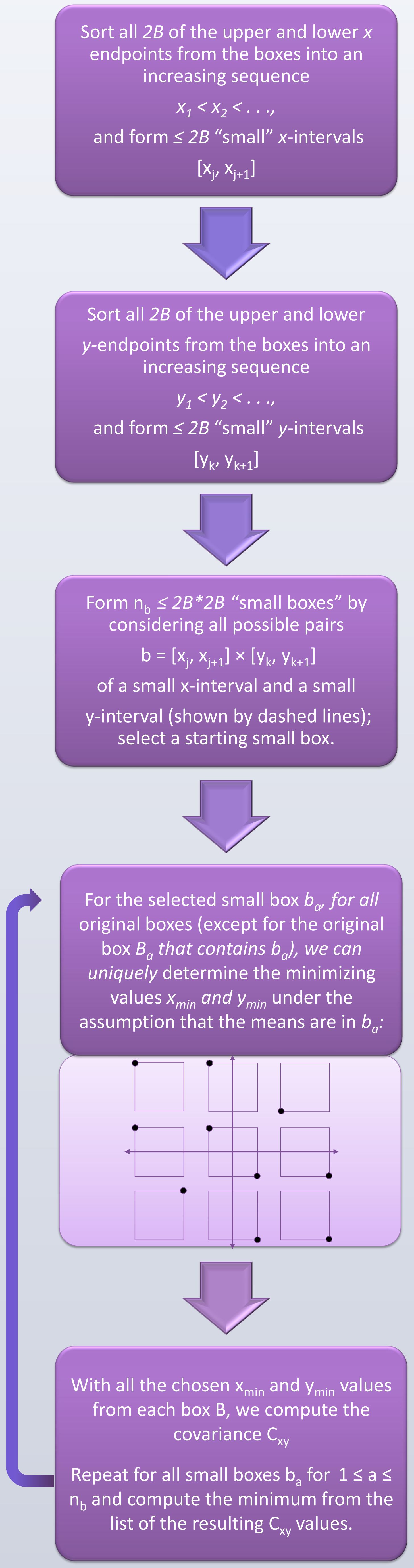
Formulation of the Problem

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.

Under the new algorithm, instead of beginning with a series of x and y intervals, we must begin with a series of B non-overlapping boxes, each with its own pair of x and y intervals.

New Algorithms

Covariance: | **Computational Time:**



Sort all *2B* of the upper and lower *x* endpoints from the boxes into an increasing sequence $x_1 < x_2 < \ldots,$ and form ≤ *2B* "small" *x*-intervals $[x_j, x_{j+1}]$ — **O(B log(B))**

Sort all *2B* of the upper and lower *y*-endpoints from the boxes into an increasing sequence $y_1 < y_2 < \ldots,$ and form ≤ *2B* "small" *y*-intervals $[y_k, y_{k+1}]$ — **O(B log(B))**

Form $n_b \le 2B*2B$ "small boxes" by considering all possible pairs $b = [x_j, x_{j+1}] \times [y_k, y_{k+1}]$ of a small x-interval and a small y-interval (shown by dashed lines); select a starting small box. — **$n_b = O(B^2)$**

For the selected small box b_a, *for all* original boxes (except for the original box B_a *that contains* b_a*), we can uniquely* determine the minimizing values x_{min} *and* y_{min} under the assumption that the means are in b_a: — **$n_b * B = O(B^3)$**

With all the chosen x_{min} and y_{min} values from each box B, we compute the covariance C_{xy}

Repeat for all small boxes b_a for $1 \le a \le n_b$ and compute the minimum from the list of the resulting C_{xy} values. — **$O(n_b) = O(B^2)$**

Final Computational Time: $O(B^3)$

Correlation: Algorithm

The same correlation algorithm as was previously known, with a minor modification, can be used for the general privacy-protecting interval data. The main difference is that instead of considering records one by one, we consider boxes one by one, and we process all the data points corresponding to each box in a single step.

Correlation: Computational Time

The computational time for the previous algorithm was bounded by $O(n^5)$. From a practical standpoint, this would be difficult because the number of records n is usually quite large.

In the privacy-motivated case with specific boxes, we have ≤ 4B vertices, where B is the number of different boxes. Thus, the total number of quadruples of vertices is $O(B^4)$. Once the quadruple is fixed, then, within each box b_a, we select the same optimizing values x_{max} and y_{max} (or x_{min} and y_{min}) for all the records from this box. Thus, we need to perform only a finite number of computations within each box. 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 $O(B^5)$ is still large, but since the number of boxes B is much smaller than the number of records n, this number of steps is much smaller than $O(n^5)$ – and thus, more realistic.

Conclusions

What we found. We developed feasible (polynomial-time) algorithms for computing statistical characteristics such as covariance and correlation based on a general privacy-protected interval data.

Practical applications. The newly developed algorithms can be used to better find dependence between different quantities in statistical databases. This is extremely important in biomedical applications, where it is necessary to find out which factors contribute to a disease and which factors affect the efficiency of different cures. Such dependences are also important in processing census data, where it is important to uncover correlations between, e.g., geographical and social characteristics– for example, how income depends on the geographical location.

Ideas for future research. First, it is desirable to see if faster algorithms are possible, especially for computing correlation for which our B5 algorithm is still somewhat slow. Second, it would be great to develop similar algorithms for computing other statistical characteristics, e.g., robust characteristics of correlation which are used for non-Gaussian data. Third, it is important to extend our algorithms to situations of hierarchical estimation, when we first, e.g., process the data within a given county and then combine the results for all counties within a state.

Acknowledgements

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, via the REU program.

It was also partly supported 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.

My sincerest thanks to all those who mentored me, especially to Dr. Vladik Kreinovich, Dr. Luc Longpre, Ali Jalal-Kamali, Victor Talavera, and Dr. Hein Nguyen.

References

1. A. Jalal-Kamali et al., "Estimating Covariance for Privacy Case under Interval (and Fuzzy) Uncertainty", Proceedings of the World Conference on Soft Computing, San Francisco, CA, May 23-26, 2011.
2. A. Jalal-Kamali et al., "Estimating Correlation under Interval Uncertainty", Mechanical Systems and Signal Processing, 37, 43–53, 2013.
3. G. Xiang et al., "Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics", Proc. 2013 IEEE Series of Symposia on Computational Intelligence.