Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics: Fuzzy-Motivated Approach

G. Xiang, S. Ferson, L. Ginzburg
Applied Biomathematics
contact email gxiang@sigmaxi.net

L. Longpré, E. Mayorga, O. Kosheleva
University of Texas at El Paso
contact email olgak@utep.edu

1. Need to Preserve Privacy

- To better serve customers, it is important to know as much as possible about them.
- Customers are often reluctant to share information, since this information can be used against them.
- For example, age can be used by companies to (unlawfully) discriminate against older job applicants.
- It is thus important to preserve privacy when storing customer data.
- To maintain privacy, we divide the space of all possible combinations of values $x = (x_1, \ldots, x_n)$ into boxes

$$B = [\widetilde{x}_1 - \Delta_1(x), \widetilde{x}_1 + \Delta_1(x)] \times \ldots \times [\widetilde{x}_n - \Delta_n(x), \widetilde{x}_n + \Delta_n(x)] \,.$$

- For each record, instead of storing the actual values x_i, we only store the label of the box B containing x.

2. k-Anonymity and ℓ-Diversity

- For each record, instead of storing the actual values x_i, we only store the label of the box B containing x.
- To avoid further loss of privacy, it is important to make sure that location in a box does not identify a person.
- This is usually achieved by requiring that for some fixed integer k, each box contains at least k records.
- This is called *k-anonymity*.
- It is also not good if all records within a box have the same value of an i-th quantity x_i.
- It is thus required that for some integer ℓ, each box should contain at least ℓ different values of each x_i.
- This is called *ℓ-diversity*.

3. Statistical Data Processing

- *Given:* data points $x^{(p)} = \left(x_1^{(p)}, \ldots, x_n^{(p)}\right)$, $1 \le p \le N$.
- We need to estimate several characteristics:
- The mean is estimated as $E_i = \dfrac{1}{N} \cdot \sum\limits_{p=1}^{N} x_i^{(p)}$.
- The covariance $C_{ij} = \dfrac{1}{N} \cdot \sum\limits_{p=1}^{N} \left(x_i^{(p)} - E_i\right) \cdot \left(x_j^{(p)} - E_j\right)$.
- The variance $V_i = \dfrac{1}{N-1} \cdot \sum\limits_{p=1}^{N} \left(x_i^{(p)} - E_i\right)^2$.
- The correlation is estimated as $\rho_{ij} = \dfrac{C_{ij}}{\sigma_i \cdot \sigma_j}$.

4. In Statistical Data Processing, Privacy Leads to Uncertainty

- To maintain privacy, we replace each numerical value $x_i^{(p)}$ with the corresponding interval.
- Different values from these intervals lead, in general, to different values of the statistical characteristics.
- Hence, for each characteristic, we get a whole interval of possible values.
- If this interval is too wide, the resulting range is useless (e.g., for correlation, the interval $[-1, 1]$ is useless).
- It is therefore desirable to select,
  - among all possible subdivisions into boxes which preserve k-anonymity (and ℓ-diversity),
  - the one which leads to the narrowest intervals for the desired statistical characteristic.

5. Uncertainty Caused by Subdivision into Boxes

- To minimize uncertainty, we select the smallest boxes.
- Hence, each box B should have exactly k records.
- For each $x_i^{(p)}$, we know the interval $\left[\widetilde{x}_i^{(p)} - \Delta_i^{(p)}, \widetilde{x}_i^{(p)} + \Delta_i^{(p)}\right]$, so $\left|\Delta x_i^{(p)}\right| \le \Delta_i^{(p)}$ for $\Delta x_k^{(p)} \stackrel{\text{def}}{=} x_k^{(p)} - \widetilde{x}_k^{(p)}$.
- Here, $C = C\left(\widetilde{x}_1^{(1)} + \Delta x_1^{(1)}, \widetilde{x}_2^{(1)} + \Delta x_2^{(1)}, \ldots, \widetilde{x}_n^{(N)} + \Delta x_n^{(N)}\right)$.
- When we have many records, boxes are small, so we can use a linear approximation:

$$C = \widetilde{C} + \sum_{p=1}^{N} \sum_{i=1}^{n} \frac{\partial C}{\partial x_i} \cdot \Delta x_i^{(p)}.$$

- The range of this linear expression is $\left[\widetilde{C} - \Delta, \widetilde{C} + \Delta\right]$, where $\Delta \stackrel{\text{def}}{=} k \cdot \sum\limits_{B} \sum\limits_{x \in B} \sum\limits_{i=1}^{n} \left|\frac{\partial C}{\partial x_i}\right| \cdot \Delta_i(x)$.

6. Expressions for the Partial Derivatives

- For all these characteristics C, the derivative takes the form $\dfrac{\partial C}{\partial x_i} = \dfrac{1}{N} \cdot b_i(x)$ for some expression $b_i(x)$.
- For the mean E_i, the derivative is equal to $\dfrac{\partial E_i}{\partial x_i} = \dfrac{1}{N}$.
- For the variance V_i, we have $\dfrac{\partial V_i}{\partial x_i} = \dfrac{2 \cdot (x_i - E_i)}{N}$.
- Therefore, for $\sigma_i = \sqrt{V_i}$, we get $\dfrac{\partial \sigma_i}{\partial x_i} = \dfrac{x_i - E_x}{N \cdot \sigma_i}$.
- For the covariance C_{ij}, we have $\dfrac{\partial C_{ij}}{\partial x_i} = \dfrac{x_j - E_j}{N}$.
- We also have: $\dfrac{\partial \rho_{ij}}{\partial x_i} = \dfrac{1}{N} \cdot \dfrac{(x_j - E_j) - \dfrac{C_{ij}}{\sigma_i^2} \cdot (x_i - E_i)}{\sigma_i \cdot \sigma_j}$.

7. Towards Optimal Subdivision into Boxes

- The overall expression for Δ is a sum of terms corresponding to different points.
- To minimize Δ, we must, for each point, minimize the corresponding term $\sum\limits_{i=1}^{n} \left|\dfrac{\partial C}{\partial x_i}\right| \cdot \Delta_i(x)$.
- The only constraint on the values $\Delta_i(x)$ is that the corresponding box should contain exactly k points.
- The number of points can be obtained by multiplying the data density $\rho(x)$ by the box volume $\prod\limits_{i=1}^{n} (2\Delta_i(x))$.
- The data density can be estimated based on the data.
- So, we minimize the expression $\sum\limits_{i=1}^{n} a_i(x) \cdot \Delta_i(x)$ under the constraint $\rho(x) \cdot 2^n \cdot \prod\limits_{i=1}^{n} \Delta_i(x) = k$.

8. (Asymptotically) Optimal Subdivision into Boxes (Case of k-Anonymity)

- The Lagrange multiplier technique leads to $\Delta_i(x) = \dfrac{c(x)}{a_i(x)}$, for some $c(x)$.

- From the constraint, we get $c(x) = \dfrac{1}{2} \cdot \sqrt[n]{\dfrac{k}{\rho(x)} \cdot \prod\limits_{j=1}^{n} a_j(x)}$.

- This means that around each point x, we need to select the box with half-widths

$$\Delta_i(x) = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \frac{\sqrt[n]{\prod\limits_{j=1}^{n} a_j(x)}}{a_i(x)}.$$

- The resulting accuracy is equal to $\Delta = \dfrac{n}{N} \cdot \sum\limits_{x} c(x)$, where the sum is taken over all N data points x.

9. We Need to Dismiss Rare Points

- In many practical situations, we have rare points, for which the smallest box containing k of them is huge.
- Such a big-size box will contribute a large amount of uncertainty to Δ; so we should dismiss such rare points.
- The privacy-related uncertainty is $\frac{n}{\#S}\cdot\sum\limits_{x\in S} c(x)$, where S is the set of remaining points.
- The statistical accuracy reduces to $\frac{A}{\sqrt{\#(S)}}$.
- Minimizing the sum $\frac{n}{\#(S)}\cdot\sum\limits_{x\in S} c(x)+\frac{A}{\sqrt{\#(S)}}$ leads to selecting all x with $c(x)\le c_0$, where c_0 minimizes

$$\frac{n}{\#\{x:c(x)\le c_0\}}\cdot\sum_{x:c(x)\le c_0} c(x)+\frac{A}{\sqrt{\#\{x:c(x)\le c_0\}}}.$$

10. Examples

- For estimating the mean E_i, we have $a_i(x) = \text{const}$ and thus, $c(x) = \text{const} \cdot \dfrac{1}{\sqrt[n]{\rho(x)}}$.
- So, dismissing points with $c(x) > c_0$ is equivalent to dismissing all the points with $\rho(x) < \rho_0$ (for some ρ_0).
- For computing covariance C_{ij}, the derivative is proportional to $x_i - E_i$.
- Thus, the values $a_i(x)$ are proportional to $|x_i - E_i|$.
- So, the upper threshold c_0 on $c(x)$ is equivalent to the lower threshold on the ratio $\dfrac{\rho(x)}{|x_i - E_i| \cdot |x_j - E_j|}$.
- Hence, we can also use points x with small $\rho(x)$, provided that if x_i or x_j is close to the corresponding mean.
- Using extra points x improves accuracy.

11. How to Also Take into Account ℓ-Diversity

- Within each box, for each variable x_i, there should be $\geq \ell$ different values of x_i.
- Different usually means that $|x_i - x'_i| \geq \varepsilon_i$ for some threshold ε_i.
- Thus, ℓ different values means that $2\Delta_i(x) \geq \ell \cdot \varepsilon_i$.
- To use this additional constraint, we first compute the values $\Delta_i(x)$ as before.
- If $2\Delta_i(x) \geq \ell \cdot \varepsilon_i$ for all i, we select $\Delta_i(x)$.
- Otherwise, we sort the quantities by $a_i(x) \cdot \varepsilon_i$:

$$a_1(x) \cdot \varepsilon_1 \geq a_2(x) \cdot \varepsilon_2 \geq \ldots \geq a_n(x) \cdot \varepsilon_n.$$

12. How to Take into Account ℓ-Diversity (cont-d)

- *Reminder:* We sort the quantities by $a_i(x)\cdot\varepsilon_i$:

$$a_1(x)\cdot\varepsilon_1 \ge a_2(x)\cdot\varepsilon_2 \ge \ldots \ge a_n(x)\cdot\varepsilon_n.$$

- Then, for each t from 1 to n, we compute

$$c_t = \frac{1}{2}\cdot\left(\frac{k\cdot\prod\limits_{i=t+1}^{n} a_i(x)}{\rho(x)\cdot\ell^t\cdot\prod\limits_{i=1}^{t}\varepsilon_i}\right)^{1/(n-t)}.$$

- For each t, if $\dfrac{2c_t}{\ell}\ge a_{t+1}(x)\cdot\varepsilon_{t+1}$, we compute

$$\Delta(t)\stackrel{\text{def}}{=}\frac{1}{2}\cdot\ell\cdot\sum_{i=1}^{t} a_i(x)\cdot\varepsilon_i+(n-t)\cdot c_t.$$

- We select t_m for which $\Delta(t)$ is the smallest, and take $\Delta_i(x)=\dfrac{1}{2}\cdot\ell\cdot\varepsilon_i$ for $i\le t_m$, $\Delta_i(x)=\dfrac{c_{t_m}}{a_i(x)}$ for $i>t_m$.

13. Fuzzy-Motivated Idea

- To improve the accuracy of the resulting estimate, we ignored some data points while keeping other data points.
- In other words, we used a crisp separation between:
  - data points that we keep and
  - data points that we ignore.
- Fuzzy logic has taught us that in many cases, it is beneficial to use a "fuzzy" separation.
- Specifically, we assign a weight $w(x) \geq 0$ to each data point so that $\sum w(x) = 1$.
- We then use weighted estimates:

$$E_i = \sum_x w(x) \cdot x_i, \quad \sigma_i^2 = \sum_x w(x) \cdot (x_i - E_i)^2.$$

$$C_{ij} = \sum_x w(x) \cdot (x_i - E_i) \cdot (x_j - E_j), \quad \rho_{ij} = \frac{C_{ij}}{\sigma_i \cdot \sigma_j}.$$

14. Optimization Problem

- Our objective is to find the weights $w(x)$ for which the resulting uncertainty is the smallest possible.
- For privacy-motivated uncertainty, the corresponding derivatives $\dfrac{\partial C}{\partial x_i}$ are proportional to the weight $w(x)$.
- As a result, for the overall privacy-motivated uncertainty, we get the expression $n \cdot \sum\limits_x w(x) \cdot c(x)$.
- The variance of an estimate $E_i = \sum w(x) \cdot x_i$ is the sum of the variances: $\sim \sum w^2(x)$.
- Thus, the standard deviation is $\sim \sqrt{\sum\limits_x w^2(x)}$.
- Problem: $n \cdot \sum\limits_x w(x) \cdot c(x) + A \cdot \sqrt{\sum\limits_x w^2(x)} \to \min$ under the constraints $\sum\limits_x w(x) = 1$ and $w(x) \geq 0$.

15. Iterative Algorithm for Computing the Auxiliary Parameter λ

- On each iteration, we first compute the total numbers $\widetilde{N}$ of points x for which $n \cdot c(x) < \lambda_k$.
- Then, we compute the sums $\sum_x c(x)$ and $\sum_x c^2(x)$ over all such points.
- Based on these values, we find λ_{k+1} from the equation

$$\widetilde{N} \cdot \lambda^2 - 2\lambda \cdot n \cdot \sum_x c(x) + n^2 \cdot \sum_x c^2(x) - A^2 = 0.$$

- Here, the sums are over all x for which $n \cdot c(x) < \lambda$.
- We stop iterations when the process converges, i.e., when $\lambda_{k+1} = \lambda_k$.
- In the process of computing λ, we have computed the values $\widetilde{N}$ and $\sum\limits_{x: n \cdot c(x) < \lambda} c(x)$.

16. Computing Optimal Weights $w(x)$

- We have computed:
  - λ,
  - $\widetilde{N} = \#\{x : n \cdot c(x) < \lambda_k\}$, and
  - $\sum\limits_{x : n \cdot c(x) < \lambda} c(x)$.
- Then, we compute

$$K = \frac{1}{\widetilde{N} \cdot \lambda - \sum\limits_{x} c(x)}.$$

- The optimal weights can now be computed as follows:

$$w(x) = \max(K \cdot (\lambda - c(x)), 0).$$

17. Acknowledgment

Support for this project was provided:

- by the National Institutes of Health (NIH),
- through a Small Business Innovation Research grant 1R43TR000173-01 to Applied Biomathematics.