A Latent Class Modeling Approach for Differentially Private Synthetic Data for Contingency Tables

Andrés Felipe Barrientos

Assistant Professor
Department of Statistics
Florida State University

Joint work with Michelle P. Nixon, Aleksandra Slavković, and Jerry P. Reiter.

2021 FCSM conference
November 2021

Outline

1. Data privacy
2. Differentially private modeling approach
3. Illustrations with 2016 ACS data
4. Concluding remarks

Privacy and data sharing

- Agencies and companies often seek to share their data.
- Protection of individuals' private information is a must.
- Traditional strategies: disclosure control methods [Hundepool et al., 2012] or releasing synthetic data [Rubin, 1993].
- In recent years, agencies are looking for methods that provide formally quantifiable privacy guarantees, e.g., those that rely on differential privacy.

Problem setup

- Confidential dataset $\boldsymbol{X} = \{X_i = (X_{1i}, \ldots, X_{pi})\}_{i=1}^n$, where X_{ij} is categorical.
- Assume that the agency is willing to release summaries of $\boldsymbol{X}$ denoted by $M(\boldsymbol{X}) = (M_1(\boldsymbol{X}), \ldots, M_T(\boldsymbol{X}))$.
- The goal is to generate a synthetic version of $\boldsymbol{X}$ using $M(\boldsymbol{X})$ and a formally private mechanism.

Illustration with ACS PUMS

- We selected a subset of 10,000 individuals from the 2016 one-year ACS PUMS.
- Each $M_t(\boldsymbol{X}),\ t = 1, \ldots, 10$, denotes a two-way marginal table.

	Age	
Citizenship	0	1
0	11	596
1	443	8950

	Race	
Citizenship	0	1
0	299	308
1	1731	7662

	Sex	
Citizenship	0	1
0	273	334
1	4505	4888

	Income	
Citizenship	0	1
0	294	313
1	2916	6477

	Race	
Age	0	1
0	110	344
1	1920	7626

	Sex	
Age	0	1
0	239	215
1	4539	5007

	Income	
Age	0	1
0	445	9
1	2765	6781

	Sex	
Race	0	1
0	945	1085
1	3833	4137

	Income	
Race	0	1
0	827	1203
1	2382	5587

	Income	
Sex	0	1
0	1281	3497
1	1929	3293

Differential privacy

- Differential privacy is the best known formal privacy framework in use.
- $\mathcal{M}(\boldsymbol{X})$ is a randomized version of $M(\boldsymbol{X})$.

Definition

ϵ-Differential Privacy [Dwork et al, 2006]: A randomized mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if for all data sets $\boldsymbol{X}$ and $\boldsymbol{X}'$ differing on at most one row, and $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{M})$,

$$\frac{\Pr[\mathcal{M}(\boldsymbol{X}) \in S | \boldsymbol{X}]}{\Pr[\mathcal{M}(\boldsymbol{X}') \in S | \boldsymbol{X}']} \leq \exp(\epsilon)\,.$$

Differentially private summary statistics

- $\mathcal{M}(\boldsymbol{X}) = (\mathcal{M}_1(\boldsymbol{X}), \ldots, \mathcal{M}_T(\boldsymbol{X}))$ is a randomized version of $M(\boldsymbol{X}) = (M_1(\boldsymbol{X}), \ldots, M_T(\boldsymbol{X}))$.

Theorem

Geometric Mechanism ***[Ghosh et. al, 2012]****: For $M_t(\boldsymbol{X}) : \mathcal{D} \to \mathbb{Z}^{d_t}$, the mechanism $\mathcal{M}_t$ that adds independently drawn noise from a* two-sided-Geom$(exp\{\frac{-\epsilon_t}{\Delta M_t}\})$ *distribution to each of the d_t terms of $M_t(\boldsymbol{X})$ satisfies ϵ_t-differential privacy.*

- Sensitivity $\Delta M_t = \sup_{\boldsymbol{X},\boldsymbol{X}'} \|M_t(\boldsymbol{X}) - M_t(\boldsymbol{X}')\|_1$.

Illustration with ACS PUMS

- Sequential composition [Mcsherry, 2009]: If each $\mathcal{M}_t$ provides ϵ_t-differential privacy. The sequence of $\mathcal{M}(\boldsymbol{X}) = (\mathcal{M}_1(\boldsymbol{X}), \ldots, \mathcal{M}_T(\boldsymbol{X}))$ provides $(\epsilon = \sum_t \epsilon_t)$-differential privacy. We can use $\epsilon_t = \epsilon / T$.

	Age	
Citizenship	0	1
0	11	596
1	443	8950

	Race	
Citizenship	0	1
0	299	308
1	1731	7662

	Sex	
Citizenship	0	1
0	273	334
1	4505	4888

	Income	
Citizenship	0	1
0	294	313
1	2916	6477

	Race	
Age	0	1
0	110	344
1	1920	7626

	Sex	
Age	0	1
0	239	215
1	4539	5007

	Income	
Age	0	1
0	445	9
1	2765	6781

	Sex	
Race	0	1
0	945	1085
1	3833	4137

	Income	
Race	0	1
0	827	1203
1	2382	5587

	Income	
Sex	0	1
0	1281	3497
1	1929	3293

Bayesian modeling approach

- The released summary statistic is of the form

$$\mathcal{M}(\boldsymbol{X}) = (M_1(\boldsymbol{X}) + \varepsilon_1, \ldots, M_T(\boldsymbol{X}) + \varepsilon_T).$$

- Some counts based on $\mathcal{M}(\boldsymbol{X})$ will not necessary match.

- Ideal modeling approach:

$$\mathcal{M}_t(\boldsymbol{X}) | M_t(\boldsymbol{X}) \overset{ind}{\sim} \text{two-sided-Geom}_{d_t}\left(M_t(\boldsymbol{X}), exp\left\{\frac{-\epsilon}{\Delta M_t T}\right\}\right),$$
$$M(\boldsymbol{X}) = (M_1(\boldsymbol{X}), \ldots, M_T(\boldsymbol{X}))|\boldsymbol{\theta} \sim p_M(\cdot|\boldsymbol{\theta}),$$
$$\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}.$$

- It is not easy to characterize $p_M(\cdot|\boldsymbol{\theta})$.

- We know that $M_t(\boldsymbol{X})|\boldsymbol{\theta} \sim \text{Multinomial}_{r_t}(n, P_t(\boldsymbol{\theta}))$.

Bayesian modeling approach using composite likelihood methods

- Proposed modeling approach:

$$\mathcal{M}_t(\boldsymbol{X})|M_t(\boldsymbol{X}) \overset{ind}{\sim} \text{two-sided-Geom}_{d_t}\left(M_t(\boldsymbol{X}), exp\left\{\frac{-\epsilon}{\Delta M_t T}\right\}\right),$$

$$M_t(\boldsymbol{X})|\boldsymbol{\theta} \overset{ind}{\sim} \text{Multinomial}_{d_t}(n, P_t(\boldsymbol{\theta})),\ t = 1, \ldots, T,$$

$$\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}.$$

- Notice that the probabilities $P_1(\boldsymbol{\theta}), \ldots, P_T(\boldsymbol{\theta})$ are related.

- We can define $P_t(\boldsymbol{\theta})$ by specifying a model for $\boldsymbol{X}|\boldsymbol{\theta}$.

Illustration with ACS PUMS

$M_1(\boldsymbol{X}) =$

	Age	
Citizenship	0	1
0	11	596
1	443	8950

$$P_1(\boldsymbol{\theta}) = \begin{pmatrix} p_{1,(0,0)} & p_{1,(0,1)} \\ p_{1,(1,0)} & p_{1,(1,1)} \end{pmatrix}$$

$M_2(\boldsymbol{X}) =$

	Race	
Citizenship	0	1
0	299	308
1	1731	7662

$$P_2(\boldsymbol{\theta}) = \begin{pmatrix} p_{2,(0,0)} & p_{2,(0,1)} \\ p_{2,(1,0)} & p_{2,(1,1)} \end{pmatrix}$$

- Coherence: $p_{1,(1,0)} + p_{1,(1,1)} = p_{2,(1,0)} + p_{2,(1,1)}$
- We define $P_t(\boldsymbol{\theta})$ by specifying a model for $\boldsymbol{X}|\boldsymbol{\theta}$.

Modeling $\boldsymbol{X}|\boldsymbol{\theta}$

- We use the following mixture model [Dunson and Xing 2009]:

$$
\begin{aligned}
X_{ij}|z_i, \{\Psi_h^{(j)}\}_{h=1}^{\infty} &\overset{ind}{\sim} \mathit{Multinomial}\{1, \Psi_{z_i 1}^{(j)}, \ldots, \Psi_{z_i d_j}^{(j)}\}, \\
z_i|\{\pi_h\}_{h=1}^{\infty} &\overset{ind}{\sim} \mathit{Discrete}\{(1, \ldots, \infty), (\pi_1, \ldots, \pi_\infty)\}, \\
\pi_h &= V_h \prod_{l<h}(1 - V_l), \quad V_h \sim \beta(1, \alpha), \\
\Psi_h^{(j)} &\sim \mathit{Dirichlet}(a_{j1}, \ldots, a_{jd_j}),
\end{aligned}
$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\pi}_k = \{\pi_h\}_{h=1}^{k},\ \boldsymbol{\Psi}_k = \{\Psi_h^{(j)}\}_{h=1,j=1}^{k,p}\right)$.

Defining $P_1(\boldsymbol{\theta}), \ldots, P_T(\boldsymbol{\theta})$

- If $M_1(\boldsymbol{X})$ is the contingency table of the first two variables, then

$$P_1(\boldsymbol{\theta}) = \begin{pmatrix} p_{1,(0,0)} & p_{1,(0,1)} \\ p_{1,(1,0)} & p_{1,(1,1)} \end{pmatrix}$$

where, e.g.,

$$p_{1,(0,0)} = Pr(X_{\cdot 1} = 0, X_{\cdot 2} = 0|\boldsymbol{\theta}) = \sum_{h=1}^{k} \pi_h \Psi_{h0}^{(1)} \Psi_{h0}^{(2)} \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{l=0}^{1} \Psi_{hi}^{(3)} \Psi_{hj}^{(4)} \Psi_{hk}^{(5)}.$$

Bayesian modeling approach and inference

- Proposed approach:

$$
\begin{aligned}
\mathcal{M}_t(\boldsymbol{X})|M_t(\boldsymbol{X}) &\overset{ind}{\sim} \text{two-sided-Geom}_{d_t}\left(M_t(\boldsymbol{X}), exp\left\{\frac{-\epsilon}{\Delta M_t T}\right\}\right), \\
M_t(\boldsymbol{X})|\boldsymbol{\theta} &\overset{ind}{\sim} \text{Multinomial}_{d_t}(n, P_t(\boldsymbol{\theta})),\ t = 1, \ldots, T, \\
\boldsymbol{\theta} &\sim p_{\boldsymbol{\theta}}.
\end{aligned}
$$

- We use MCMC algorithms to sample from $\boldsymbol{\theta}|\mathcal{M}(\boldsymbol{X})$.

- Inferences are performed using $(P_1(\boldsymbol{\theta}), \ldots, P_T(\boldsymbol{\theta}))\,|\mathcal{M}(\boldsymbol{X})$.

Bayesian modeling approach and inference

- Instead of using $M(\boldsymbol{X})|\mathcal{M}(\boldsymbol{X})$, we use $M(\boldsymbol{X}^S)|\mathcal{M}(\boldsymbol{X})$.

- To make inferences about the confidential summary, we use

$$Pr(X_{(n+1)1} = c_1, \ldots, X_{(n+1)p} = c_p|\mathcal{M}(\boldsymbol{X})) =$$
$$\int Pr(X_{(n+1)1} = c_1, \ldots, X_{(n+1)p} = c_p|\boldsymbol{\theta})Pr(\boldsymbol{\theta}|\mathcal{M}(\boldsymbol{X}))d\boldsymbol{\theta}$$

to generate synthetic datasets $\boldsymbol{X}^S$ and induce a distribution via $\boldsymbol{X}^S \mapsto M(\boldsymbol{X}^S)$.

Illustrations with ACS PUMS

- We selected a subset of 10,000 individuals from the 2016 one-year ACS PUMS.
- Each $M_t(\boldsymbol{X}),\ t = 1, \ldots, 10$, denotes a two-way marginal table.

	Age	
Citizenship	0	1
0	11	596
1	443	8950

	Race	
Citizenship	0	1
0	299	308
1	1731	7662

	Sex	
Citizenship	0	1
0	273	334
1	4505	4888

	Income	
Citizenship	0	1
0	294	313
1	2916	6477

	Race	
Age	0	1
0	110	344
1	1920	7626

	Sex	
Age	0	1
0	239	215
1	4539	5007

	Income	
Age	0	1
0	445	9
1	2765	6781

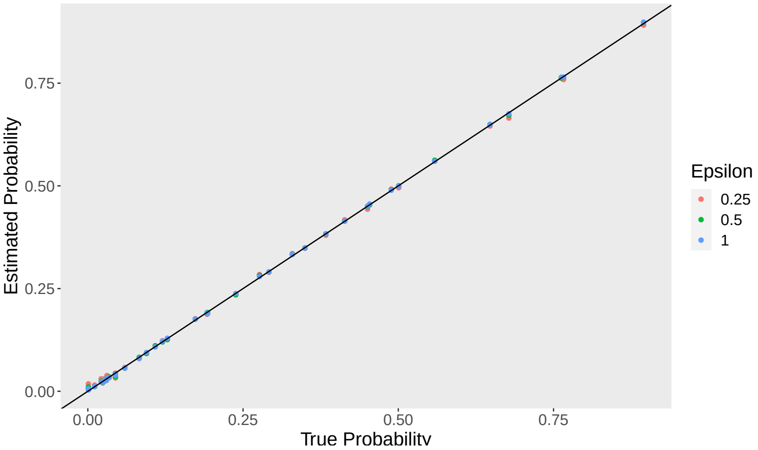
	Sex	
Race	0	1
0	945	1085
1	3833	4137

	Income	
Race	0	1
0	827	1203
1	2382	5587

	Income	
Sex	0	1
0	1281	3497
1	1929	3293

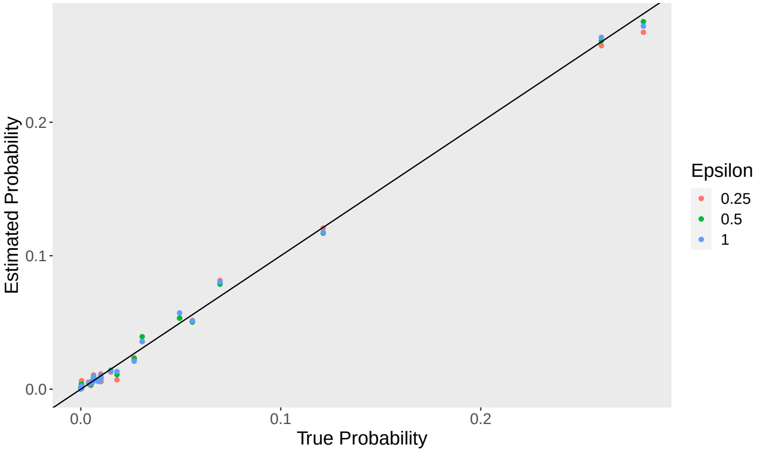
Illustrations with ACS PUMS

- True versus estimated two-way marginal tables.



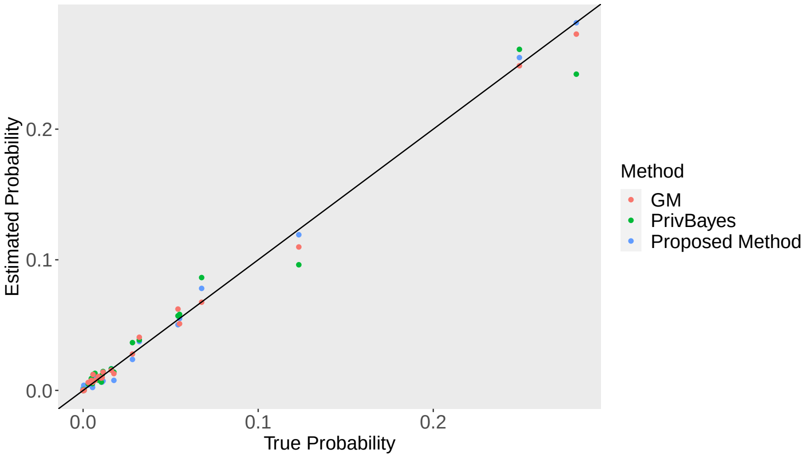
Illustrations with ACS PUMS

- True versus estimated full table.



Comparisons with existing methods

- True versus estimated full table ($\epsilon = 0.5$).



Concluding remarks

- We present a novel method to create differentially private synthetic data for contingency tables based on marginal counts.
- The simulation results indicate that our approach preserves the summaries.
- The proposed approach is complementary to existing releasing mechanisms.
- Our general strategy can be extended to more complex data structures.

Thank you!