Synthetic population generation for nested data using differentially private posteriors

Hang J. Kim[1] Terrance Savitsky[2] *Matt Williams*[3]
Monika Hu[4]

[1]University of Cincinnati (hang.kim@uc.edu)

[2]Bureau of Labor Statistics (Savitsky.Terrance@bls.gov)

[3]RTI International (mrwilliams@rti.org)

[4]Vassar College (jihu@vassar.edu)

FCSM
Oct 25, 2023

Outline

1 Motivating Examples

2 Review of Differential Privacy

3 Extensions to Nested Privacy

Motivating Data Structures

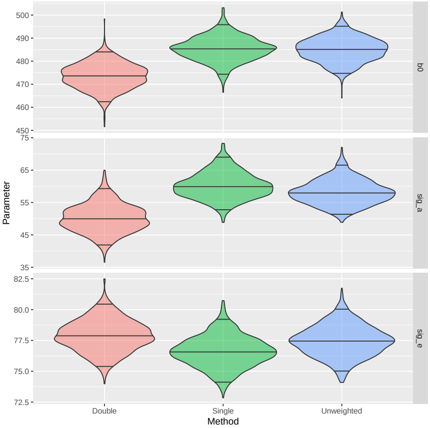
- Data Set with Nested Entities
  - Students >> Teachers (class)
  - Employees >> Owners (business)
  - Patients >> Doctors (hospitals)
- Entities in each level may have disclosure concerns
  - poor performance
  - sensitive responses
  - competitive advantage
  - risk of regulatory intervention

Motivating Models

- Variance Decomposition
  - Attribute % of variability to group vs. individual factors

$$y_{ig} = \mu_g + \epsilon_i$$
$$\mu_g \sim N(\nu, \tau^2)$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- PISA 2000: Science Scores (US)
  - Average Score (top)
  - Between Class Variation (mid)
  - Individual Variation (bottom)
  - Different estimation methods (columns/colors)



Outline

1 Motivating Examples

2 Review of Differential Privacy

3 Extensions to Nested Privacy

Differential privacy (Dwork et al., 2006)

Let $D \in \mathbb{R}^{n \times k}$ be a database in input space $\mathcal{D}$. Let $\mathcal{M}$ be a randomized mechanism such that $\mathcal{M}() : \mathbb{R}^{n \times k} \to O$. Then $\mathcal{M}$ is ϵ-differentially private if

$$\frac{Pr[\mathcal{M}(D) \in O]}{Pr[\mathcal{M}(D') \in O]} \leq \exp(\epsilon),$$

for all possible outputs $O = Range(\mathcal{M})$ under all possible pairs of *neighboring* datasets $D, D^{'} \in \mathcal{D}$

- An output statistic f on database D: $f(D)$
- Global sensitivity $\Delta = \sup_{D,D' \in \mathcal{D}:\ \delta(D,D')=1} \mid f(D) - f(D^{'}) \mid$
- Definition of neighborhood - difference of an individual. We will focus on Leave One Out (LOO).
- Laplace Mechanism for additive noise, scaled to be proportional to Δ_G/ϵ with $\epsilon-$DP guarantee

DP or not DP?

Why DP?

- Guarantee is global over all databases and provable.
- DP is property of a probabilistic mechanism. Plausible deniability.
- No explicit assumptions about intruder behaviors or knowledge
- Additivity of privacy guarantee across releases based on worst case sensitivity (not averaging). Same privacy 'currency' for very different data uses (tables, model output, public use file creation, etc).
- Privacy parameter ϵ is a finite resource that needs to be budgeted.

DP or not DP?

Why not DP?

- Worst case sensitivity often ∞. Mechanisms with $\epsilon < \infty$ can be challenging to prove (or implement).
- In practice assumptions of bounded data space not correct (e.g. value of sales for a company).
- Supremum (maximum) criteria often severely injures data utility.
- Privacy is not really a single dimension ϵ and is context-specific.

The Exponential Mechanism

- Wasserman and Zhou (2010); Zhang et al. (2016); Snoke and Slavkovic (2018) propose the Exponential Mechanism (EM) to generate synthetic data with DP properties.
- The EM generates samples from

$$\hat{\theta} \propto \exp\left(u(x,\theta)\right)\pi\left(\theta \mid \gamma\right), \tag{1}$$

  where $u(x,\theta)$ is a utility function with bound Δ_u, $\pi(\theta \mid \gamma)$ is the "base" distribution to ensure a proper density function (Zhang et al., 2016; McSherry and Talwar, 2007).
- A single sample $\hat{\boldsymbol{\theta}}_j$ has a DP guarantee of $\epsilon \leq 2\Delta_u$.
- Using a globally bounded $u(x,\theta)$ is difficult. Rejection and Metropolis Hastings sampling do not scale well with the dimension of θ.

The Posterior Mechanism and the Exponential Mechanism

- Consider the log-likelihood function as the utility function, i.e. $u(x,\theta) = \log\left(\prod_{i=1}^{n} \pi\left(x_i \mid \theta\right)\right)$ and the prior distribution $\pi(\theta \mid \gamma)$ is the base measure.
- Posterior Mechanism is an instantiation of the Exponential Mechanism

$$\exp\left(\log\left(\prod_{i=1}^{n} \pi\left(x_i \mid \theta\right)\right)\right)\pi\left(\theta \mid \gamma\right) = \left(\prod_{i=1}^{n} \pi\left(x_i \mid \theta\right)\right)\pi\left(\theta \mid \gamma\right)$$

- Sampling from a Posterior is well researched and supported!

Generalizing the (Exponential and) Posterior Mechanisms

- To reduce $\epsilon < 2\Delta_u$, modify the utility function $u(\mathrm{x}, \theta)$.
- Rescale it: $u^*(\mathrm{x}, \theta) = \frac{\epsilon}{2\Delta_u} u(\mathrm{x}, \theta)$ if $\Delta_u < \infty$. (See McSherry and Talwar, 2007, among many others).
- Scalar-weighted pseudo-likelihood (posterior)

$$\exp\left(\frac{\epsilon \log\left(\prod_{i=1}^{n} \pi\left(x_i \mid \theta\right)\right)}{2\Delta}\right) \xi\left(\theta \mid \gamma\right) = \left(\prod_{i=1}^{n} \pi\left(x_i \mid \theta\right)^{\frac{\epsilon}{2\Delta}}\right) \pi\left(\theta \mid \gamma\right)$$

Pseudo Posterior Mechanism

- Savitsky et al. (2019) utilize record-indexed weights, $\alpha \in (0, 1]^n$
- To downweight likelihood contributions with high disclosure risk

$$\xi^{\boldsymbol{\alpha}}(\theta \mid \mathrm{x}, \gamma) \propto \left[\prod_{i=1}^{n} \pi(x_i \mid \theta)^{\alpha_i}\right] \pi(\theta \mid \gamma)$$

- $\alpha_i \propto 1/\sup_{\theta \in \Theta} \mid f_\theta(x_i) \mid$
- Allows surgical downweighting of high risk records
- α_i induces an anti-informative prior
- Ensures $\Delta_\alpha < \infty$
- Expected to better preserve real data distribution for any target privacy budget, ϵ

Outline

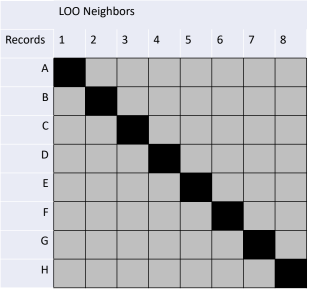
1 Motivating Examples

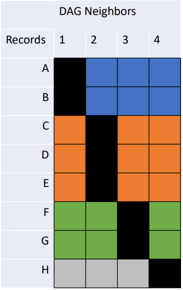
2 Review of Differential Privacy

3 Extensions to Nested Privacy

Extending the Neighborhood

- Leave-one-group-out (LOGO) or delete-a-group (DAG)
- Neighbors D and D' differ by an entire group (school, hospital, business)
- Global sensitivity $\Delta_G = \sup_{D,D' \in \mathcal{D}:\ \delta(D,D')=1_G} \mid f(D) - f(D') \mid$





Extending to Latent Variables

- Hierarchical Model

$$[y_{gi}|\mu_g] \sim N(\mu_g, \sigma^2),$$
$$[\mu_g|\nu] \sim N(\nu, \tau^2)$$

with data as response y_{gi} and group indicator 1_g and latent group mean μ_g.

- The utility function is then (the log of) the integrated likelihood

$$u_G(x, \theta) = \sum_{g=1}^{G} \log \int \left(\prod_{i=1}^{n_g} f(y_{gi}|\mu_g, \sigma^2) \right) f(\mu_g|\nu, \tau^2)\ d\mu_g$$

- We assess the LOO and DAG sensitivities of $u_G(x, \theta)$ to measure the individual and group level DP bounds ϵ.

Extending the Weighting Approach

Where should we insert the weights?

$$u_G^A(x,\theta) = \sum_{g=1}^{G} \alpha_g \log \int \left(\prod_{i=1}^{n_g} f(y_{gi}|\mu_g, \sigma^2) \right) f(\mu_g|\nu, \tau^2)\ d\mu_g$$

$$u_G^B(x,\theta) = \sum_{g=1}^{G} \log \int \left[\left(\prod_{i=1}^{n_g} f(y_{gi}|\mu_g, \sigma^2) \right) f(\mu_g|\nu, \tau^2) \right]^{\alpha_g} d\mu_g$$

$$u_G^C(x,\theta) = \sum_{g=1}^{G} \log \int \left[\left(\prod_{i=1}^{n_g} f(y_{gi}|\mu_g, \sigma^2)^{\alpha_{gi}} \right) f(\mu_g|\nu, \tau^2) \right]^{\alpha_g} d\mu_g$$

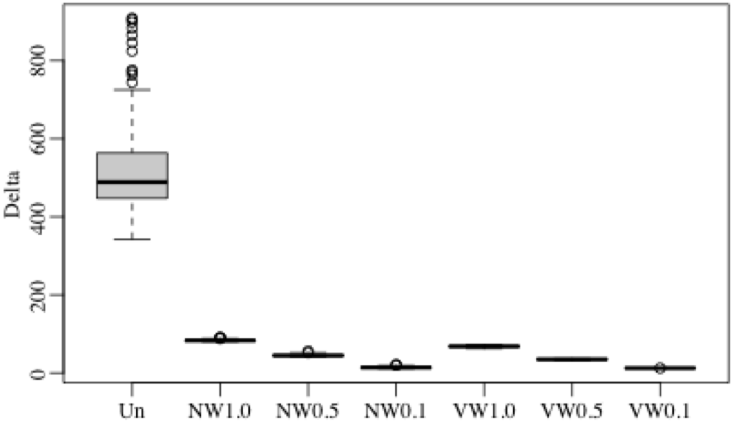
Option (A) requires us to have analytic integration for estimation. (B) and (C) allow for data augmentation approaches for estimation. (C) allows for individual-level tuning.

Preliminary Simulation Results

- $G = 100$ groups with $n_g = 50$ individuals
- $p_g, p_{gi} \in [0, 1]$: approximate risk measures based on DAG and LOO
- All down-weighting schemes reduce group level sensitivity (Delta)
- Additional (vector) down-weighting of groups α_g after vector down-weighting of individuals α_{gi} - little gain in privacy or utility

Name	alpha_g	alpha_gi	sig2	tau2	Delta_1Q	Delta_Med	Delta_3Q
Un	1	1	1.001	4.088	447.6	488.7	563.3
NW1.0	1	(1-p_gi)	0.654	4.173	82.1	83.2	84.6
NW0.5	1	.5(1-p_gi)	0.67	4.157	43.3	44.6	46.7
NW0.1	1	.1(1-p_gi)	0.841	3.992	13	13.9	15.4
VW1.0	(1-p_g)	(1-p_gi)	0.63	3.756	66.5	68.3	69.5
VW0.5	(1-p_g)	.5(1-p_gi)	0.663	3.36	33.9	34.9	35.5
VW0.1	(1-p_g)	.1(1-p_gi)	0.909	2.893	12.3	12.3	12.3

Group Level Sensitivity



Challenges and Future Work

- Calculating the sensitivities for LOO and DAG require tractable integrals. Numeric and other approximations might be possible.
- Data augmentation can still be used for parameter generation.
- Most of the gains in privacy seem to come from the individual weights α_{gi} with little additional gains from α_g. We are investigating this more.
- Group level privacy ϵ are naturally larger than individual level. While an acceptable individual level ϵ might be in $[0.1, 10]$ - its not clear what the target ϵ for groups should be $[n_g/10, 10n_g] = [5, 500]$?

References I

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006), Calibrating noise to sensitivity in private data analysis, *in* 'Proceedings of the Third Conference on Theory of Cryptography', TCC'06, Springer-Verlag, Berlin, Heidelberg, pp. 265–284.
URL: *http://dx.doi.org/10.1007/11681878*$_1$4

McSherry, M. and Talwar, K. (2007), Mechanism design via differential privacy, *in* 'Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science', pp. 94–103.

Savitsky, T. D., Williams, M. R. and Hu, J. (2019), 'Bayesian Pseudo Posterior Mechanism under Differential Privacy', *arXiv e-prints* p. arXiv:1909.11796.

Snoke, J. and Slavkovic, A. (2018), pMSE mechanism: Differentially private synthetic data with maximal distributional similarity., *in* J. Domingo-Ferrer and F. Montes, eds, 'Privacy in Statistical Databases', Vol. 11126 of *Lecture Notes in Computer Science*, Springer, pp. 138–159.

Wasserman, L. and Zhou, S. (2010), 'A statistical framework for differential privacy', *Journal of the American Statistical Association* **105**, 375–389.

Zhang, Z., Rubinstein, B. I. P. and Dimitrakakis, C. (2016), On the differential privacy of Bayesian inference, *in* 'Proceedings of the 30th AAAI Conference on Artificial Intelligence', AAAI, pp. 2365–2371.