Bayesian Pseudo Posterior Mechanism under Asymptotic Differential Privacy

Terrance D. Savitsky [1] Matthew R. Williams [2]
Monika (Jingchen) Hu [3]

[1] U.S. Bureau of Labor Statistics (Office of Survey Methods Research)

[2]NSF (National Center for Science and Engineering Statistics)

[3]Vassar College (Mathematics and Statistics Department)

JSM - 2024 August 7, 2024

Outline

Differential Privacy

Synthetic Data

Pseudo Posterior Mechanism
- Construction of Risk-based Weights
- Achieving Global DP Guarantee

Simulation Study to Illustrate Asymptotic Contraction of Lipschitz

Applications to the Consumer Expenditure Surveys

Re-weighting Strategy

Posterior Predictive Distribution

- Distribution of replicate data $\mathbf{x}^*$ given (|) observed data, $\mathbf{x}$
- Used for imputation or to generate synthetic (or "fake") data
- Model-based synthetic data smooths real data distribution
- Smoothing encodes privacy while preserving essential properties of observed data distribution
- User allowed unlimited 'interrogation' of synthetic data

$$\pi\left(\mathbf{x}^*|\mathbf{x}\right) = \int \prod_{i=1}^{n} \pi\left(x_i^*|\boldsymbol{\theta}\right) \times \xi\left(\boldsymbol{\theta}|\mathbf{x}\right) d\boldsymbol{\theta} \approx \frac{1}{S} \sum_{s=1}^{S} \prod_{i=1}^{n} \pi\left(x_i^*|\boldsymbol{\theta}_s\right)$$

- Where $s = 1, \ldots, S$ indexes the number of MCMC draws of $\boldsymbol{\theta}$
- We call this model (distribution) for generating data a synthesizer

$\mathcal{M}$ = Posterior distribution

$$\xi(\theta \mid \boldsymbol{x}) \propto \prod_{i=1}^{n} \pi(x_i \mid \theta) \times \xi(\theta)$$

- Sensitivity of $\xi(\theta \mid \boldsymbol{x})$ based on $f_\theta(\boldsymbol{x}) = \log \prod_{i=1}^{n} \pi(x_i \mid \theta)$.
- $\Delta = \sup_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}^n : \delta(\boldsymbol{x},\boldsymbol{x}')=1} \sup_{\theta \in \Theta} |f_\theta(\boldsymbol{x}) - f_\theta(\boldsymbol{x}')|$
- Generate a single dataset under $\epsilon = 2\Delta$ given a draw of $\boldsymbol{\theta}$
- Each posterior draw with $\epsilon = 2\Delta$ produces one synthetic $\boldsymbol{x}^*$

Differential Privacy under $\mathcal{M} = \xi(\theta \mid \mathbf{x})$

$$\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}^n : \delta(\mathbf{x},\mathbf{x}')=1} \sup_{B \in \beta_\Theta} \frac{\xi(B \mid \mathbf{x})}{\xi(B \mid \mathbf{x}')} \leq e^{\epsilon},$$

- ϵ bounds the **change** in the probability measure ξ
  - from the inclusion of a **single record** $\delta(\mathbf{x}, \mathbf{x}') = 1$,
  - over **all possible outcomes**, $B \in \beta_\Theta$ – sets in the space of measurable sets of Θ.
  - over **all possible data sets** $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ of size n.

Outline

Differential Privacy

Synthetic Data

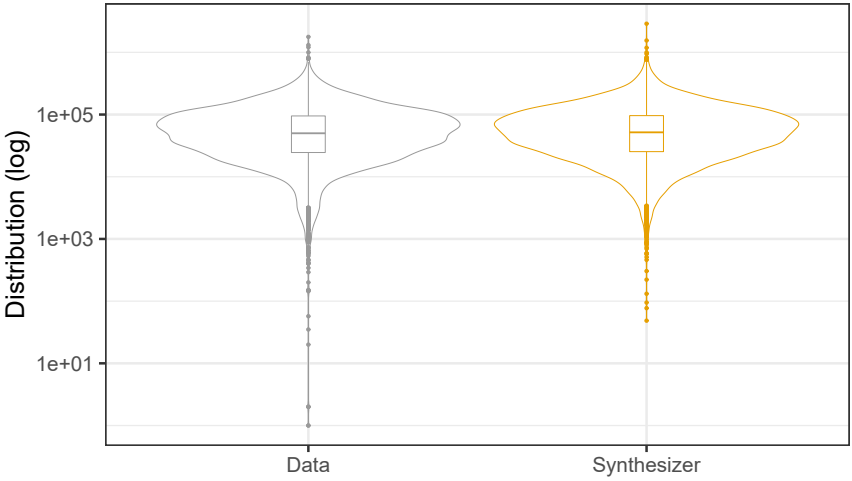
Pseudo Posterior Mechanism
- Construction of Risk-based Weights
- Achieving Global DP Guarantee

Simulation Study to Illustrate Asymptotic Contraction of Lipschitz

Applications to the Consumer Expenditure Surveys

Re-weighting Strategy

Example of a Flexible Synthesizer (Hu and Savitsky, 2019)



Outline

Differential Privacy

Synthetic Data

Pseudo Posterior Mechanism
- Construction of Risk-based Weights
- Achieving Global DP Guarantee

Simulation Study to Illustrate Asymptotic Contraction of Lipschitz

Applications to the Consumer Expenditure Surveys

Re-weighting Strategy

DP under the Posterior Mechanism

Dimitrakakis et al. (2017) establish a link between bounding the log-likelihood $f_\theta(\mathbf{x}) = \log \pi_\theta(\mathbf{x})$ and a DP bound ϵ.

- If $\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}^n : \delta(\mathbf{x},\mathbf{x}')=1} \sup_{\theta \in \Theta} |f_\theta(\mathbf{x}) - f_\theta(\mathbf{x}')| \leq \Delta$.
- Then a single draw of replicate data $\mathbf{x}^* \sim \pi(\mathbf{x}^*|\mathbf{x})$ has a DP guarantee of $\epsilon \leq 2\Delta$.

Major Limitations of the Posterior Mechanism

The results in Dimitrakakis et al. (2017) are not immediately practical.

- For many data distributions of interest (e.g. normal, exponential, Poisson, geometric), $\Delta = \infty$.
  - Directly truncating the support Θ and the domain $\mathcal{X}^n$ may work for simple distributions but is somewhat ad-hoc and scales poorly with the dimension of both
- Using a prior that induces more smoothing may reduce Δ – but requires re-estimation. This is an indirect adjustment.

Pseudo Posterior Mechanism

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

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

- $\alpha_i \propto 1/\sup_{\theta\in\Theta} |f_\theta(x_i)|$
- 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

Formulating Risk-Based Weights

How might we choose $\alpha \in (0,1]^n$?

$$\alpha_i = m\left(\sup_{\theta\in\Theta} |f_\theta(x_i)|\right), \tag{1}$$

where $f_\theta(x_i)$ is unweighted log-likelihood contribution from x_i, and $m(z)$ is a decreasing function $m : R^+ \to [0,1]$. $(m(0) = 1,\ m(\infty) = 0)$

- Riskier values downweighted more
  - α_i induces an anti-informative prior
  - $\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}^n:\delta(\mathbf{x},\mathbf{x}')=1} \sup_{\theta\in\Theta} |\alpha(\mathbf{x}) f_\theta(\mathbf{x}) - \alpha(\mathbf{x}') f_\theta(\mathbf{x}')| \le \Delta_\alpha$
  - $\Delta_\alpha \le \Delta$
- Ensures $\Delta_\alpha < \infty$
  - Produces DP budget of $\epsilon \le 2\Delta_\alpha$.
  - Control ϵ indirectly through setting of $(\alpha_1, \ldots, \alpha_n)$

Estimation Algorithm

1. Estimate unweighted $\boldsymbol{\theta}$ with model,
   $\xi(\boldsymbol{\theta}|\mathbf{x}) \propto [\prod_{i=1}^{n} \pi(x_i|\boldsymbol{\theta})] \times \pi(\boldsymbol{\theta})$

2. Compute weights, $\alpha_i = m(\sup_{\theta\in\Theta} |f_\theta(x_i)|) \propto 1/\sup_{\theta\in\Theta} |f_\theta(x_i)|$

3. Re-estimate $\boldsymbol{\theta}$ using weights, α_i in
   $\xi^{\boldsymbol{\alpha}}(\theta \mid \mathbf{x}, \gamma) \propto [\prod_{i=1}^{n} \pi(x_i \mid \theta)^{\alpha_i}] \pi(\theta \mid \gamma)$

4. Compute log-likelihood bound,
   $\sup_{\theta\in\Theta} |\alpha_i \times f_\theta(x_i)| \leq \Delta_{\alpha,x_i}$
   $\max_{x_i\in\mathbf{x}^n} \Delta_{\alpha,x_i} \leq \Delta_{\alpha,\mathbf{x}}$

5. Gives us privacy guarantee, $\epsilon \leq 2\Delta_{\alpha,\mathbf{x}}$

6. Generate synthetic data $\mathbf{x}^* \sim \pi_{\boldsymbol{\alpha}}(\mathbf{x}^*|\mathbf{x})$

Local vs. Global Privacy Guarantee

When we implement

- We have to estimate $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ based on a single data set, $\mathbf{x}$, to estimate the DP guarantee ϵ. (local DP result)
- We have to approximate the $\sup_{\theta\in\Theta} f_\theta(x_i)$ as $\max_{\theta_j, j\in 1,\ldots,J} f_{\theta_j}(x_i)$.
- Overestimate $\boldsymbol{\alpha} \rightarrow$ Underestimate ϵ

Ensuring Global DP

To justify a global DP result (bound all data sets) compared to a local DP result (bound observed data set):

- Asymptotic Discovery - For large sample sizes (n)
  - Space of plausible values Θ collapses to a point θ^*, so don't need to look at $\sup_{\theta \in \Theta}$.
  - Variation across local $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ collapses onto $\Delta_{\boldsymbol{\alpha}}$.
  - Achieves $(\epsilon, \delta)-$ pDP, where $\delta > 0$ is probability $\exists \mathbf{x} \in \mathcal{X}^n$ exceeding the ϵ bound.
    - $\delta \to 0$ at $\mathcal{O}(n^{-1/2})$.
  - Requires increasing sparsity in downweighted record contributions, which aligns with focus on isolated records as risky.
- Truncating the weights, $(\alpha_1, \ldots, \alpha_n)$
  - Set $\alpha_i^* = 0$ if $\alpha_i \times f_\theta(x_i) > \Delta_{\boldsymbol{\alpha}}$
  - Target, global threshold, $\Delta_{\boldsymbol{\alpha}}$, regulates downweighting and speed of contraction.

Outline

Differential Privacy

Synthetic Data

Pseudo Posterior Mechanism
- Construction of Risk-based Weights
- Achieving Global DP Guarantee

Simulation Study to Illustrate Asymptotic Contraction of Lipschitz

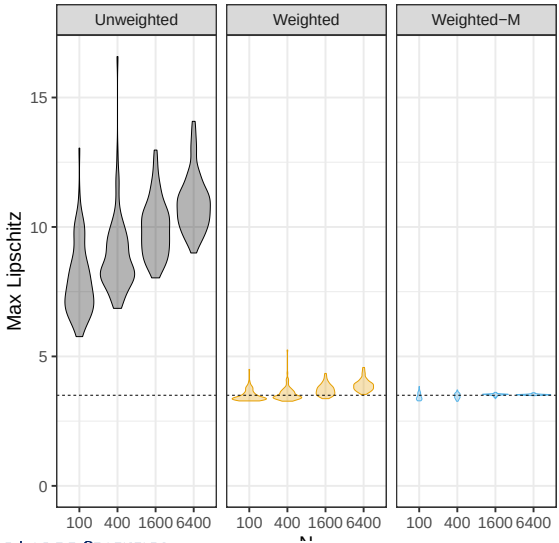
Applications to the Consumer Expenditure Surveys

Re-weighting Strategy

Monte Carlo Simulation at Increasing n

1. For sample size, $n \in \{100, 400, 1600, 6400\}$, repeat the following Monte Carlo procedure to generate a distribution of local Lipschitz bounds:
2. For $r = 1, \ldots, 100$:
   - 2.1 Generate $\mathbf{x}_r \sim \text{Pois}(\mu = 100)$, each of size n.
   - 2.2 Compute the *local* Sensitivity bound, $\Delta_{\boldsymbol{\alpha}, \mathbf{x}_r}$, for the unweighted, $\boldsymbol{\alpha}-$weighted, and $M-$ truncation-weighted pseudo posterior mechanisms.
   - 2.3 Construct the distribution of $\Delta_{\boldsymbol{\alpha}, \mathbf{x}_r}$ and note the maximum and difference between the maximum and minimum values of the distribution.
3. Assess contraction of the $\max_r \Delta_{\boldsymbol{\alpha}, \mathbf{x}_r}$ to a single (global) value and whether the minimum and maximum values collapse together.

Contraction of Weighted-M Lipschitz, $(\Delta_{\boldsymbol{\alpha},\mathbf{x}_r})$



Weighted-M Utility Performance

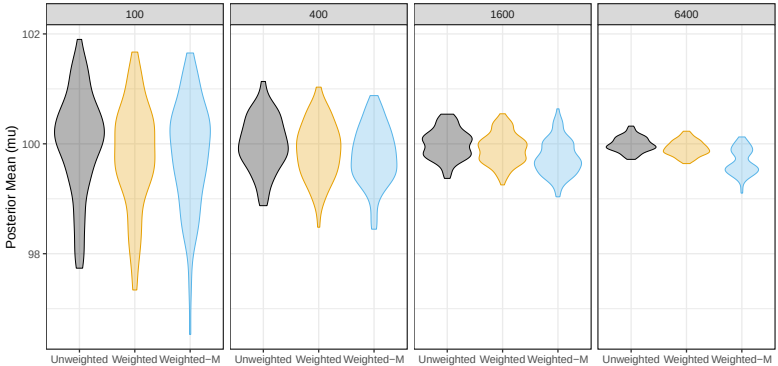


Figure: Distributions of the average of mean parameter μ for each of sample size $(100, 400, 1600, 6400)$ from $R = 100$ realizations.

Outline

Differential Privacy

Synthetic Data

Pseudo Posterior Mechanism
- Construction of Risk-based Weights
- Achieving Global DP Guarantee

Simulation Study to Illustrate Asymptotic Contraction of Lipschitz

Applications to the Consumer Expenditure Surveys

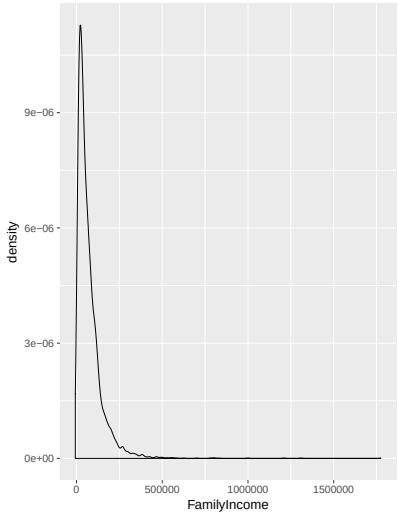
Re-weighting Strategy

Application: The Consumer Expenditure Surveys

- Provides income and expenditure patterns among consumer units (CUs)
- CE releases public-use microdata (PUMS)
- Focus on sensitive, skewed income variable – with non-sensitive variables as predictors
- From 1st QTR of 2017 survey of $n = 6208$ CUs

Variables in the CE sample

Gender
Age
Education Level
Region
Urban
Marital Status
Urban Type
CBSA
Family Size
Earner
Family Income



Comparing Synthesizer Bounds by Record ($\epsilon = 2\Delta$)

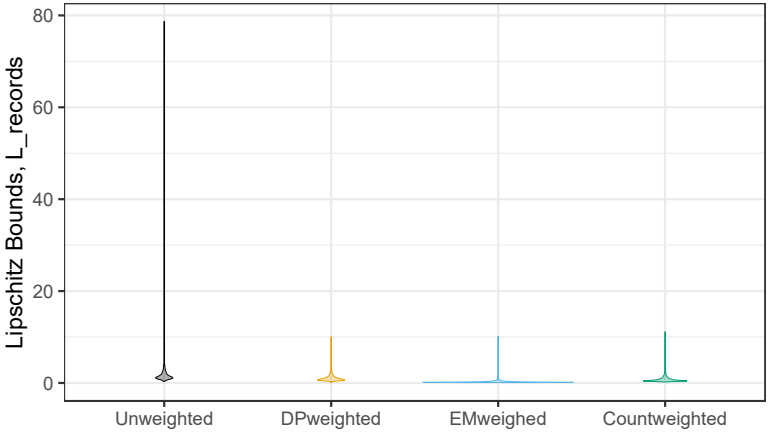
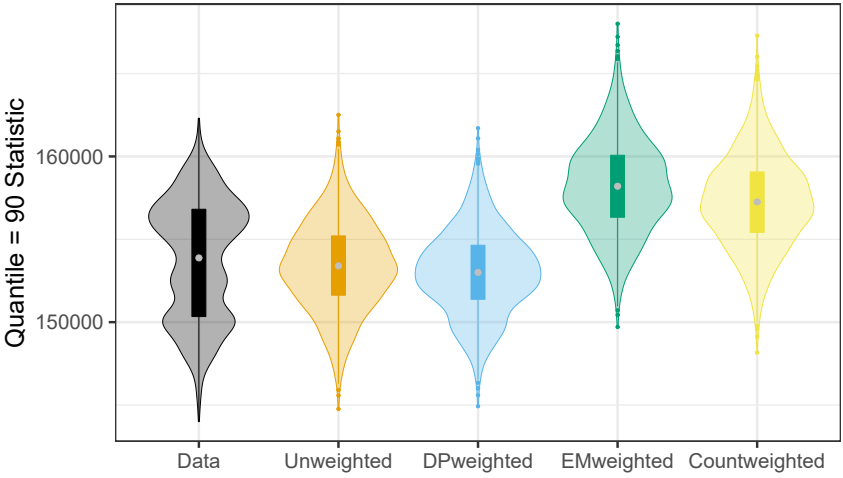


Figure: $\Delta_{Unweighted} = 78.7, \Delta_{DP} = 10.1, \Delta_{EM} = 10.2, \Delta_{Count} = 11.17.$

Comparing Synthesizer Utility (Q90) of Synthetic Data



Outline

Differential Privacy

Synthetic Data

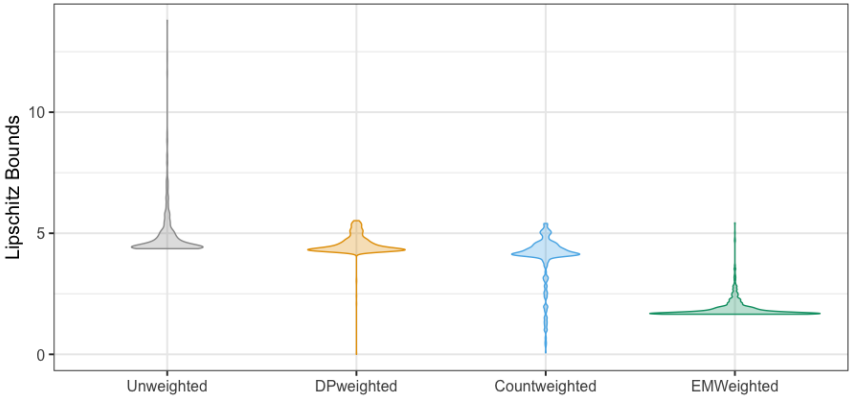
Pseudo Posterior Mechanism
- Construction of Risk-based Weights
- Achieving Global DP Guarantee

Simulation Study to Illustrate Asymptotic Contraction of Lipschitz

Applications to the Consumer Expenditure Surveys

Re-weighting Strategy

Highly Skewed Count Data



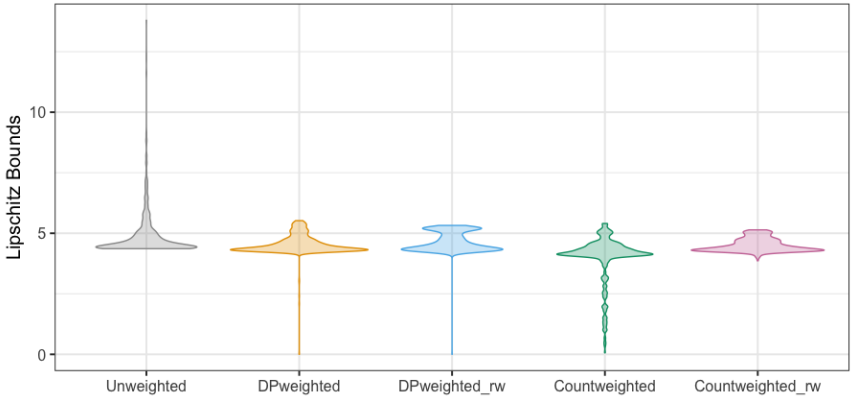
Re-weighting to compress Lipschitz Distribution

- Formal privacy guarantee is a supremum (not an average)
- There is no credit for records whose $\Delta_{\boldsymbol{\alpha},x_i} < \Delta_{\boldsymbol{\alpha},\mathbf{x}}$
- Up-weight records with $\Delta_{\boldsymbol{\alpha},x_i} < \Delta_{\boldsymbol{\alpha},\mathbf{x}}$

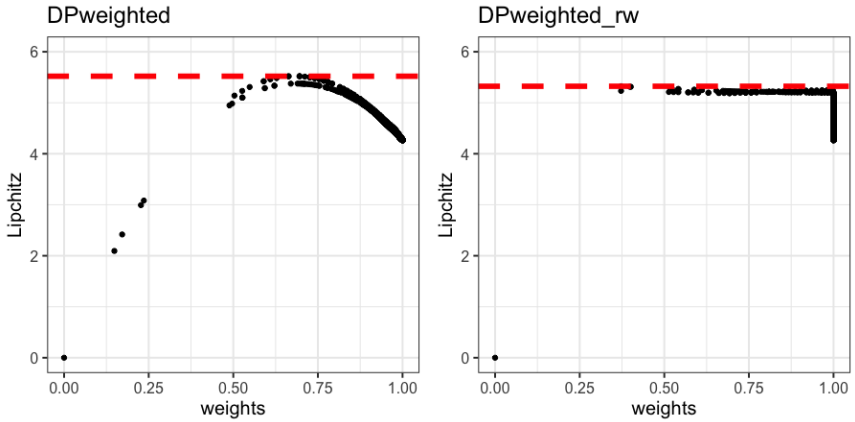
$$\alpha_i^w = k \times \alpha_i \times \frac{\Delta_{\boldsymbol{\alpha},\mathbf{x}}}{\Delta_{\boldsymbol{\alpha},x_i}} \tag{2}$$

- Compresses distribution of $\Delta_{\boldsymbol{\alpha},x_i}$
- Makes weighting scheme more efficient
- $k \leq 1$ ensures main $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$
- In the limit the weights are independent $\Delta_{\boldsymbol{\alpha},x_i}$

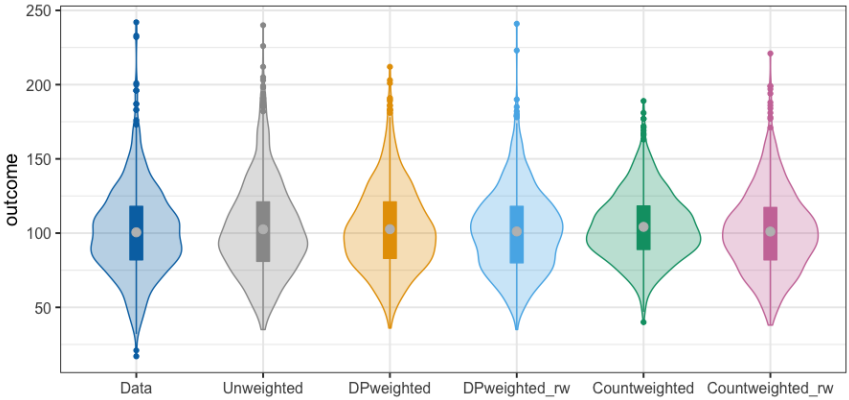
Compression of $\Delta_{\boldsymbol{\alpha},x_i}$ from Reweighting



DP Change in Weighting Efficiency



Improved Estimation of Data Distribution



References I

Dimitrakakis, C., Nelson, B., Zhang, Z., Mitrokotsa, A. and Rubinstein, B. I. P. (2017), 'Differential privacy for bayesian inference through posterior sampling', *J. Mach. Learn. Res.* **18**(1), 343–381.
URL: *http://dl.acm.org/citation.cfm?id=3122009.3122020*

Hu, J. and Savitsky, T. D. (2019), 'Bayesian Pseudo Posterior Synthesis for Data Privacy Protection', *arXiv e-prints* p. arXiv:1901.06462.

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