Bayesian Pseudo Posterior Synthesis for Data Privacy Protection

Based on works of M. Hu and T. D. Savitsky

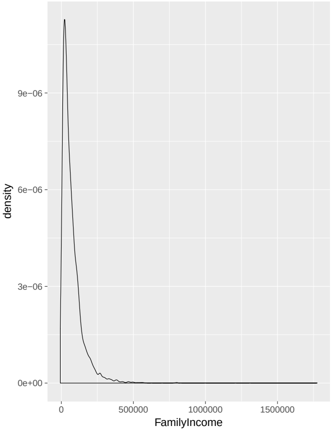
Terrance D. Savitsky

Office of Survey Methods Research

CNSTAT Workshop on Data Privacy Protection
June 7, 2019

Data: CU Income for Consumer Expenditure Survey

- Heavily right-skewed. “Bill Gates” problem



Truncated Dirichlet Process

- Flexible synthesizer to preserve data distribution.
- Smooths response values and mixes records.

$$y_i|\mathbf{x}_i, \pi_k, \boldsymbol{\beta}_k^*, \sigma_k^* \overset{\text{ind}}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}\left(y_i|\mathbf{x}_i^{'}\boldsymbol{\beta}_k^*, \sigma_k^*\right)$$

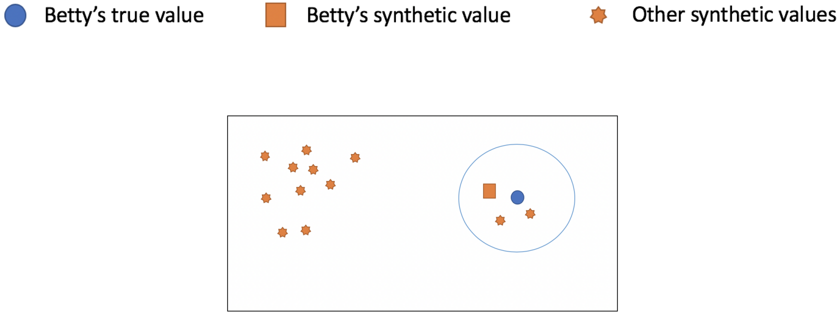
$$\pi_1, \ldots, \pi_K \sim \mathcal{D}\left(\frac{\gamma}{K}, \ldots, \frac{\gamma}{K}\right)$$

- Usual practice generate synthetic observations, $(y_1^*, \ldots, y_n^*)$

$$y_i^*|\mathbf{y} \overset{\text{ind}}{\sim} \int \left[\sum_{k=1}^{K} \pi_k \mathcal{N}\left(y_i^*|\mathbf{x}_i^{'}\boldsymbol{\beta}_k^*, \sigma_k^*\right)\right] \times \prod_{k=1}^{K} p\left(\left(\pi_k, \boldsymbol{\beta}_k^*, \sigma_k^*\right)|\mathbf{y}\right) d\left(\left(\pi_k, \boldsymbol{\beta}_k^*, \sigma_k^*\right)\right)$$

Evaluation of identification disclosure risks

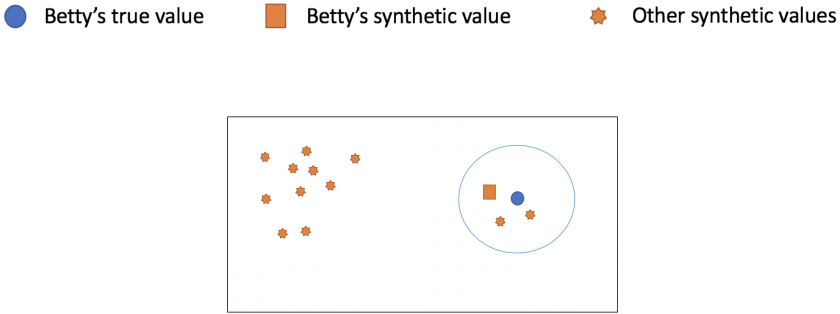
- Fewer synthetic values inside the interval/ball $\rightarrow$ the intruder has a higher probability of guessing the record of the name they seek.



Scenario 1: $IR_i = \frac{10}{13} \times 1 = \frac{10}{13}$.

Evaluation of identification disclosure risks

- Fewer synthetic values inside the interval/ball $\rightarrow$ the intruder has a higher probability of guessing the record of the name they seek.



Scenario 1: $IR_i = \frac{10}{13} \times 1 = \frac{10}{13}$.

Pseudo Posterior

- Risk-based record-indexed weights, $\alpha_i \in \{0, 1\}$

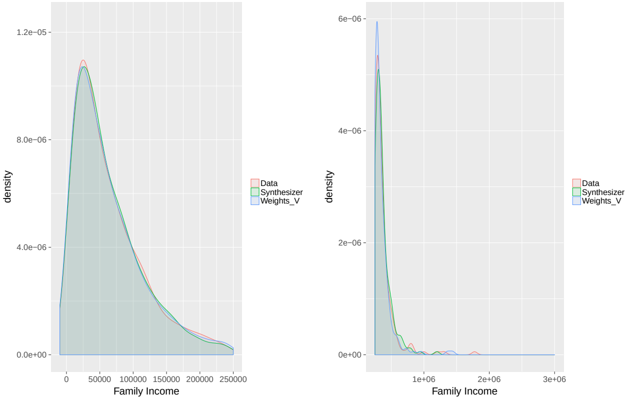
$$\alpha_i = \text{min}(0, 1 - c_i \times IR_i)$$

- For monotone constants, $(c_i) \in \max\left(0, [c_{\min} = 1.0, c_{\max} = 1.5]\right)$.
- Used to construct risk-weighted, pseudo posterior:

$$p_\alpha\left((\pi_k, \boldsymbol{\beta}^*_k, \sigma^*_k)_{k=1,\ldots,K} \mid \mathbf{y}, \theta\right) \propto \left[\prod_{i=1}^{n} p\left(y_i \mid (\pi_k, \boldsymbol{\beta}^*_k, \sigma^*_k)_{k=1}^{K}\right)^{\alpha_i}\right] \times \prod_{k=1}^{K} p\left(\pi_k, \boldsymbol{\beta}^*_k, \sigma^*_k \mid \theta\right)$$

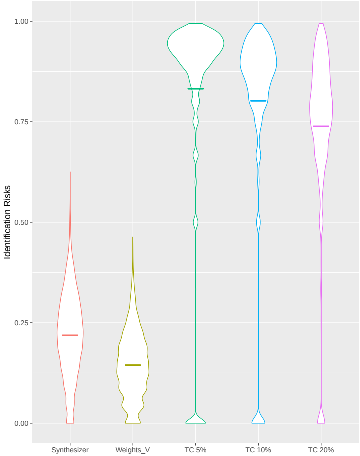
Application to CE Income: Utility

- Mass of distribution largely unaffected
- See the concentration effect in the tails
- Pulls more isolated records to the modes



CE Income: Compare Risks of Vector vs. Top-coding

- Known pattern: {gender, age, education, marital status, earner}.
- Top-coding only protects CUs with extreme incomes (see bulbs).
- Ignores other risky portions of data distribution.



CONTACT INFORMATION

Savitsky.Terrance@bls.gov