Risk-Efficient Bayesian Data Synthesis for Privacy Protection

Jingchen (Monika) Hu

Vassar College

Joint work with Terrance D. Savitsky (Bureau of Labor Statistics)
and Matthew R. Williams (National Center for Science and Engineering Statistics)

ICES VI 2021
slides available at: `http://bit.ly/ICES-VI`

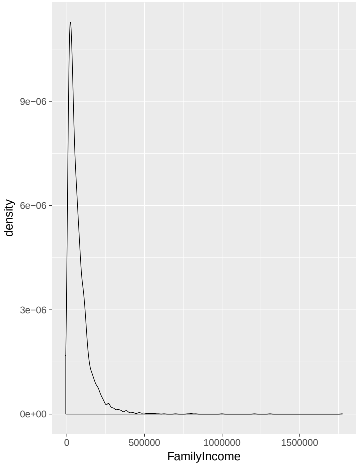
Outline

1. The problem: CE income data
2. Synthetic data and probability of identification risk
3. Risk-adjusted synthesizer
4. “Whack-a-mole” risk pop-ups and our solution
5. Summary and references

Outline

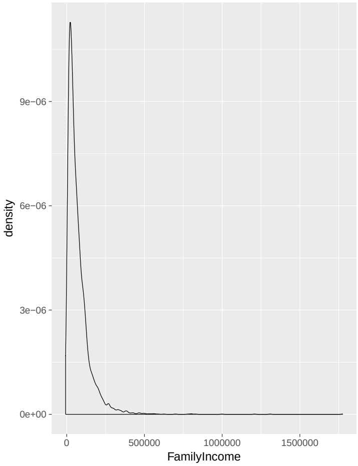
1. The problem: CE income data
2. Synthetic data and probability of identification risk
3. Risk-adjusted synthesizer
4. “Whack-a-mole” risk pop-ups and our solution
5. Summary and references

The CE data: highly skewed family income



- The Consumer Expenditure Surveys (CE).
- Family income: before tax for each consumer unit (CU).
- High risk records assumed to be in the tail, e.g., CUs with extremely high family income.
- What to do? Topcoding (statistical disclosure control).

The CE data: highly skewed family income



- The Consumer Expenditure Surveys (CE).
- Family income: before tax for each consumer unit (CU).
- High risk records assumed to be in the tail, e.g., CUs with extremely high family income.
- What to do? Topcoding (statistical disclosure control).

Outline

1. The problem: CE income data
2. Synthetic data and probability of identification risk
3. Risk-adjusted synthesizer
4. “Whack-a-mole” risk pop-ups and our solution
5. Summary and references

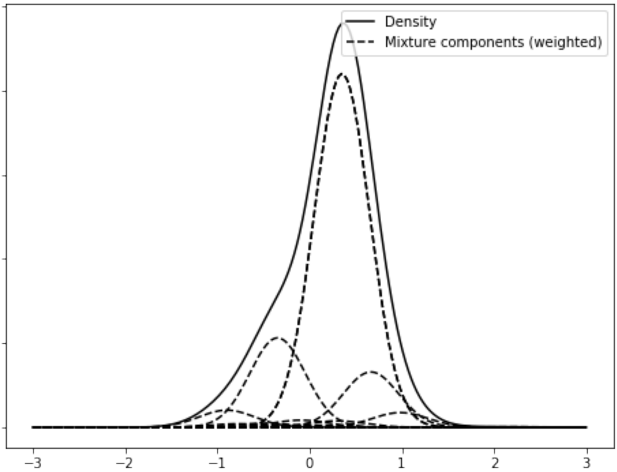
Why synthetic data?

Rubin (1993) and Little (1993) proposed the synthetic data.

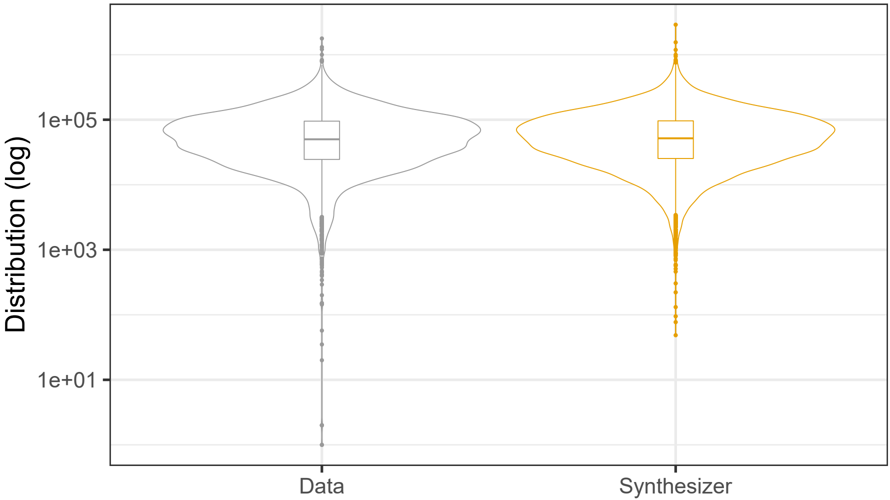
- Simulate records from statistical models that are estimated from the original confidential data.
- Balance of data utility and disclosure risks
  - preserve relationships of variables
  - low disclosure risks
- Allow data analysts to make valid inference for a wide class of analyses.

Example of a flexible synthesizer

A two-level hierarchical parametric finite mixture synthesizer



Synthesizer induces smoothing of data distribution



Risk evaluation: Intruder's knowledge and behavior

Variable	Description
Gender	Gender of the reference person; 2 categories
Age	Age of the reference person; 5 categories
Region	Region of the CU; 4 categories
Education Level	Education level of the reference person; 8 categories
Urban	Urban status of the CU; 2 categories
Marital Status	Marital status of the reference person; 5 categories
Urban Type	Urban area type of the CU; 3 categories
Family Size	Size of the CU; 11 categories
Earner	Earner status of the reference person; 2 categories
Family Income	Imputed and reported income before tax of the CU;

- A known pattern of the un-synthesized categorical variables, $\mathbf{X}_i^p \subseteq \mathbf{X}_i$, e.g. (Gender, Age, Region).
- The true value of synthesized family income y_i.
- A name or identity of interest.

Identification risks based on notion of isolation

- Define radius r of synthetic data y^* in pattern p around the truth y.
- Use percentage radius, e.g. $r = 20\%$.
  - e.g. For a CU i with $50,000 family income, the interval / ball is: [$40,000, $60,000].
- Outside of radius $\rightarrow$ isolation.
- Do this for each record y_i: all y_j^*'s in pattern p.
- Identification risk (IR) for each record is a probability.

Risk as probability of identification disclosure

Formally, for CU i with pattern p:

$$\begin{aligned} IR_i &:= \text{Pr}\,(\text{identification disclosure of } i) \\ &= \frac{\sum_{j \in M_{p,i}} \mathbb{I}\left(y_j^* \notin B(y_i, r)\right)}{|M_{p,i}|} \times T_i. \end{aligned} \quad (1)$$

- $IR_i \in [0, 1]$.
- Larger IR_i and closer to 1: higher risks.
- Smaller IR_i and closer to 0: lower risks.

Outline

1. The problem: CE income data
2. Synthetic data and probability of identification risk
3. Risk-adjusted synthesizer
4. “Whack-a-mole” risk pop-ups and our solution
5. Summary and references

A new risk-adjusted synthesizer

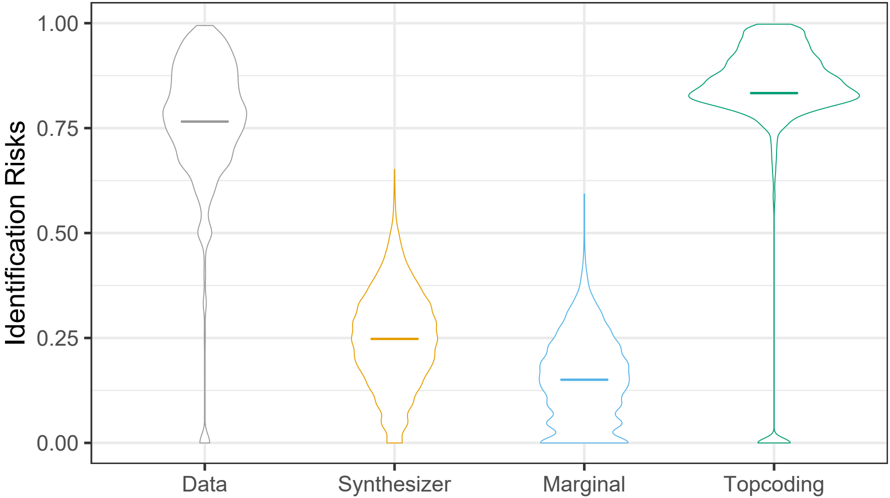
- Use weight $\alpha_i \in [0, 1]$ for CU i.
- Evaluate IR_i^c in the confidential data.
- $\alpha_i = 1 - IR_i^c$: higher $IR_i^c \rightarrow$ higher risk $\rightarrow$ lower α_i.
- Selectively downweight to defeat the likelihood principle:

$$\left[\prod_{i=1}^{n} p\left(y_i \mid \boldsymbol{\theta}\right)^{\alpha_i}\right] p\left(\boldsymbol{\theta} \mid \gamma\right). \tag{2}$$

  - $\boldsymbol{\theta}$: model parameters
  - γ: model hyperparameters

- Surgical distortion: scalar α vs vector α_i.

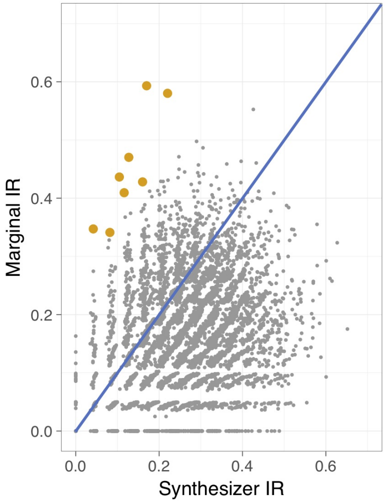
Violin plots of identification risks



Outline

1. The problem: CE income data
2. Synthetic data and probability of identification risk
3. Risk-adjusted synthesizer
4. "Whack-a-mole" risk pop-ups and our solution
5. Summary and references

Whack-a-mole issue



- Risky record value shrinkage leaves moderate risk record exposed

Pairwise weighting ties records together

- Marginal risk probability for CU i with pattern p:

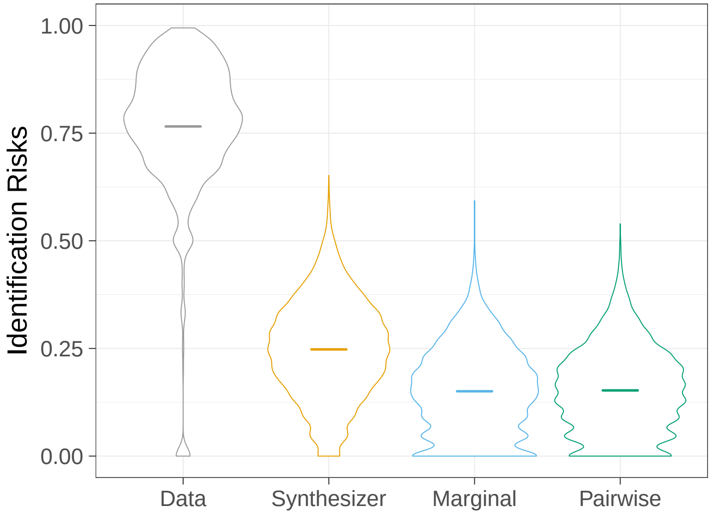
$$IR_i^c = \frac{\sum_{h \in M_{p,i}} \mathbb{I}\left(y_h \notin B(y_i, r)\right)}{|M_{p,i}|}, \; \alpha_i = 1 - IR_i^c.$$

- Pairwise risk probability for pairs of CUs (i, j) in *same* pattern p:

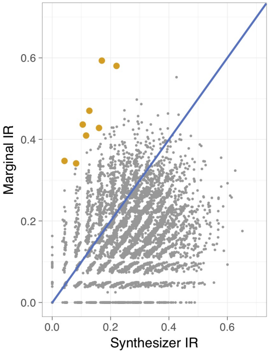
$$IR_{i,j}^c = \frac{\sum_{h \in M_{p,(i,j)}} \mathbb{I}\left(y_h \notin B(y_i, r) \cap y_h \notin B(y_j, r)\right)}{|M_{p,(i,j)}|}, \; \alpha_{i,j} = 1 - IR_{i,j}^c.$$

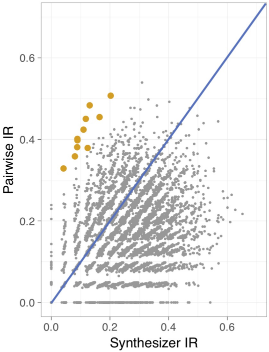
- $(\tilde{\alpha}_i)$ within each pattern constructed as dependent.
- Leave moderate-risk records more covered in the synthetic data.
- Expected to mitigate the *whack-a-mole*.

Pairwise distribution of risks is more compressed



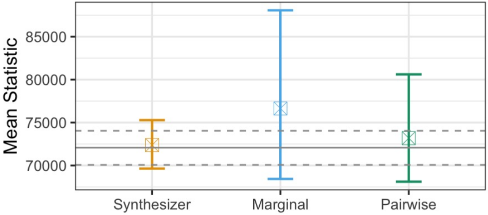
Pairwise partially resolves whack-a-mole issues

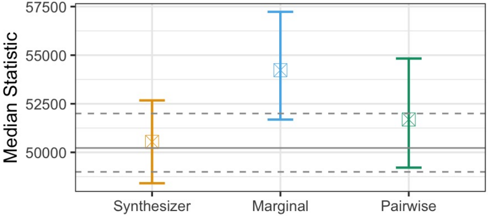




Results of utility

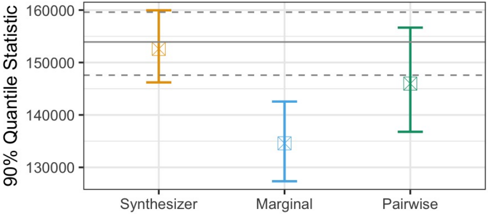
Horizontal lines: mean (solid) and 95% confidence intervals (dashed) from confidential data.

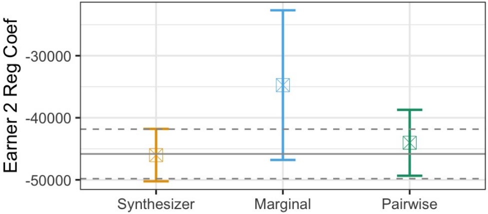




Results of utility cont’d

Horizontal lines: mean (solid) and 95% confidence intervals (dashed) from confidential data.





Outline

1. The problem: CE income data
2. Synthetic data and probability of identification risk
3. Risk-adjusted synthesizer
4. “Whack-a-mole” risk pop-ups and our solution
5. Summary and references

Summary

- A general framework to achieve desired utility-risk trade-off balance.
- Downweight scheme works for any synthesizer with high utility.
- Pairwise downweight is more risk-efficient: better control of the identification risks and little loss of utility.
- Local weight adjustments can improve utility preservation and little loss of privacy protection.
- The use of topcoding incorrectly assumes which records express high risks.

Summary

- A general framework to achieve desired utility-risk trade-off balance.
- Downweight scheme works for any synthesizer with high utility.
- Pairwise downweight is more risk-efficient: better control of the identification risks and little loss of utility.
- Local weight adjustments can improve utility preservation and little loss of privacy protection.
- The use of topcoding incorrectly assumes which records express high risks.

Summary

- A general framework to achieve desired utility-risk trade-off balance.
- Downweight scheme works for any synthesizer with high utility.
- Pairwise downweight is more risk-efficient: better control of the identification risks and little loss of utility.
- Local weight adjustments can improve utility preservation and little loss of privacy protection.
- The use of topcoding incorrectly assumes which records express high risks.

References

- An, D. and Little, R. J. A. (2007), Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A*, 170, 923-940.
- Hu, J., Savitsky, T. D. and Williams, M. R. (forthcoming), Risk-efficient Bayesian pseudo posterior data synthesis for privacy protection, *Journal of Survey Statistics and Methodology*.
- Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* 9, 407-426.
- Rubin, D. B. (1993). Discussion statistical disclosure limitation. *Journal of Official Statistics* 9, 461-468.

Example of a flexible synthesizer

Two level hierarchical parametric finite mixture synthesizer:

$$
\begin{aligned}
y_i \mid \mathbf{X}_i, z_i, \mathbf{B}^*, \boldsymbol{\sigma}^* &\sim \text{Normal}(y_i \mid \mathbf{x}_i^{'} \boldsymbol{\beta}_{z_i}^*, \sigma_{z_i}^*), && (3)\\
z_i \mid \pi &\sim \text{Multinomial}(1; \pi_1, \cdots, \pi_K), && (4)
\end{aligned}
$$

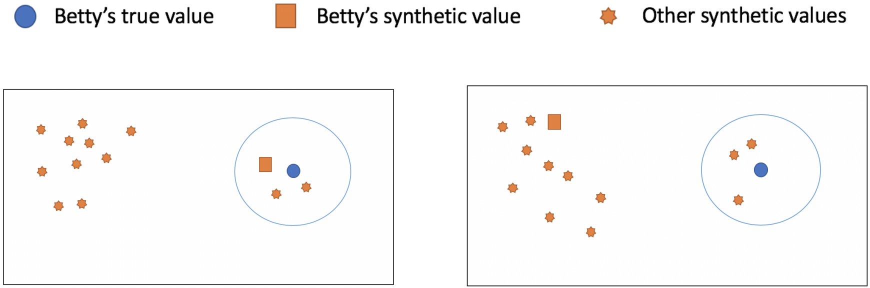
We induce sparsity in the number of clusters with,

$$
\begin{aligned}
(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right), && (5)\\
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha). && (6)
\end{aligned}
$$

Estimate the location, scale, proportion, and number of clusters

Toy example: compute probability of identification

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

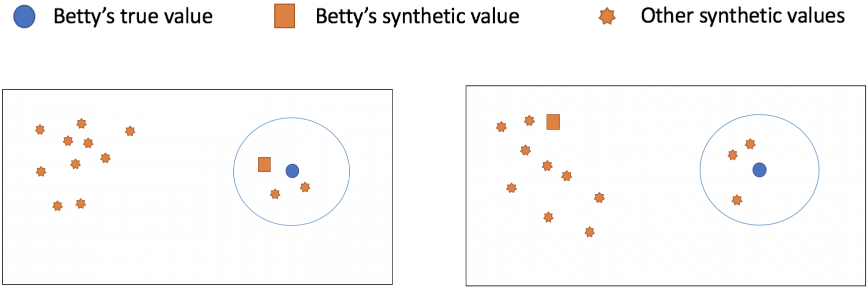


Scenario 1:
$IR_i = \frac{10}{13} \times 1 = \frac{10}{13}$.

Scenario 2:
$IR_i = \frac{10}{13} \times 0 = 0$.

Toy example: compute probability of identification

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

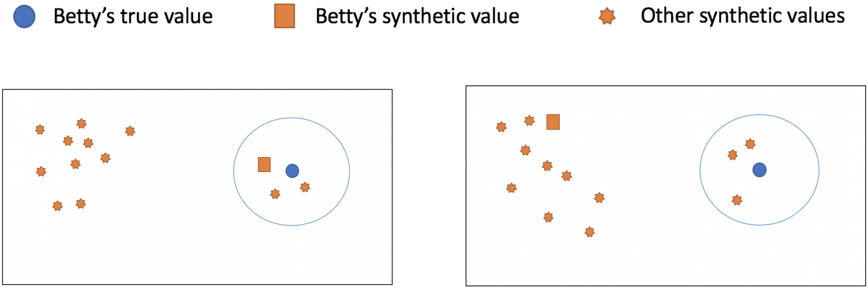


Scenario 1:
$IR_i = \frac{10}{13} \times 1 = \frac{10}{13}$.

Scenario 2:
$IR_i = \frac{10}{13} \times 0 = 0$.

Toy example: compute probability of identification

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

Scenario 2:
$IR_i = \frac{10}{13} \times 0 = 0$.

Pairwise weighting ties records together

- Pairwise risk probability for pairs of CUs (i, j) in *same* pattern p

$$IR^c_{i,j} = \frac{\sum_{\ell \in M_{p,(i,j)}} \mathbb{I}\left(y_h \notin B(y_i, r) \cap y_h \notin B(y_j, r)\right)}{|M_{p,(i,j)}|}. \quad (7)$$

$$\alpha_{i,j} = 1 - IR^c_{i,j}, \;\; \propto 1/IR^c_{i,j} \quad (8)$$

- Sum over all $\alpha_{i,j}$ for $j \neq i$, and divide by $|M_{p,i}| - 1$
- $(\tilde{\alpha}_i)$ within each pattern constructed as dependent.
- Reduce degree of shrinking of each high-risk record and leave moderate-risk records more covered in the synthetic data
- Expected to mitigate the *whack-a-mole*

$$\tilde{\alpha}_i = \frac{\sum_{j=1, j \neq i \in M_{p,i}} \alpha_{i,j}}{|M_{p,i}| - 1}. \quad (9)$$