Select synthetic microdata for survey data

Hang J. Kim

Division of Statistics and Data Science

University of Cincinnati

presented at 2024 Joint Statistical Meeting, Portland, OR

August 6, 2024

Select data synthesis approach

"Synthesize a **select set of variables** and **select records** with high disclosure risks" [from the session abstract]

1. Fully synthetic data (Rahunathan et al. 2003)

2. Partially synthetic data (Reiter 2003)
   - Often relying on a (Bayesian) conditional probability model to generate the **select** set of variables/records Y in the original data (X, Y):

$$\tilde{Y}_{\text{synt}} \sim f(\tilde{Y}|X,Y) = \int f(\tilde{Y}|X,\theta) f(\theta|X,Y) \mathrm{d}\theta \propto \int f(\tilde{Y}|X,\theta) f(Y|X,\theta) f(\theta) \mathrm{d}\theta$$

   - Using the combining rule for correct uncertainty quantification
     - under both Bayesian and repeated sampling frameworks

     $\Rightarrow$ Some deep-learning approaches (e.g., table-GAN) overlook the uncertainty quantification

Select data synthesis approach

"Synthesize a **select set of variables** and **select records** with high disclosure risks" [from the session abstract]

1. Fully synthetic data (Rahunathan et al. 2003)

2. Partially synthetic data (Reiter 2003)
   - Often relying on a (Bayesian) conditional probability model to generate the **select** set of variables/records Y in the original data (X, Y):

$$\tilde{Y}_{\text{synt}} \sim f(\tilde{Y}|X,Y) = \int f(\tilde{Y}|X,\theta) f(\theta|X,Y) \mathrm{d}\theta \propto \int f(\tilde{Y}|X,\theta) f(Y|X,\theta) f(\theta) \mathrm{d}\theta$$

   - Using the combining rule for correct uncertainty quantification
     - under both Bayesian and repeated sampling frameworks

     $\Rightarrow$ Some deep-learning approaches (e.g., table-GAN) overlook the uncertainty quantification

Select data synthesis approach

"Synthesize a **select set of variables** and **select records** with high disclosure risks" [from the session abstract]

1. Fully synthetic data (Rahunathan et al. 2003)

2. Partially synthetic data (Reiter 2003)
   - Often relying on a (Bayesian) conditional probability model to generate the **select** set of variables/records Y in the original data (X, Y):

     $$\tilde{Y}_{\text{synt}} \sim f(\tilde{Y}|X,Y) = \int f(\tilde{Y}|X,\theta) f(\theta|X,Y) \mathrm{d}\theta \propto \int f(\tilde{Y}|X,\theta) f(Y|X,\theta) f(\theta) \mathrm{d}\theta$$

   - Using the combining rule for correct uncertainty quantification
     - under both Bayesian and repeated sampling frameworks

     ⇒ Some deep-learning approaches (e.g., table-GAN) overlook the uncertainty quantification

Select data synthesis approach

"Synthesize a **select set of variables** and **select records** with high disclosure risks" [from the session abstract]

1. Fully synthetic data (Rahunathan et al. 2003)

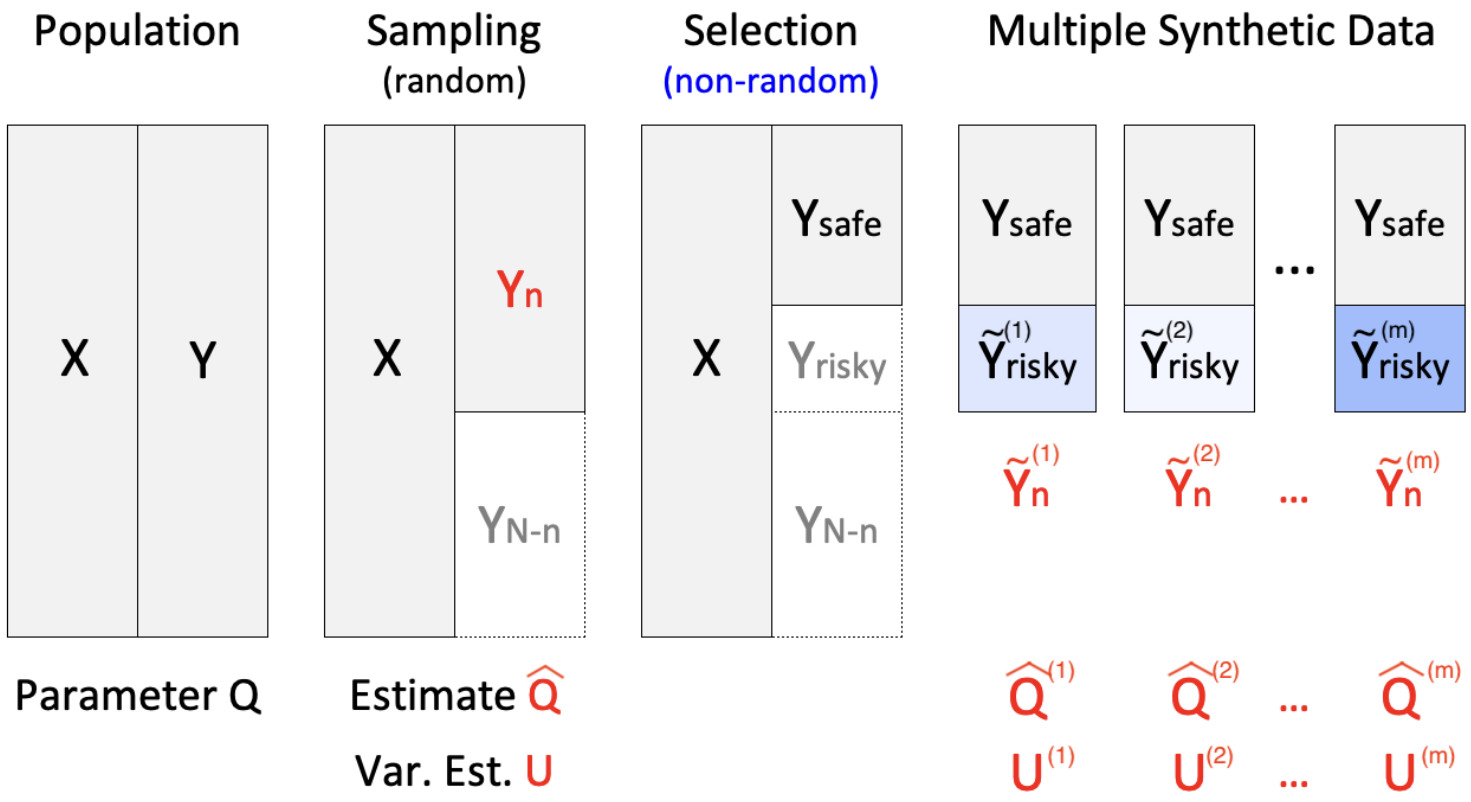
2. Partially synthetic data (Reiter 2003)
   - Often relying on a (Bayesian) conditional probability model to generate the **select** set of variables/records Y in the original data (X, Y):

     $$\tilde{Y}_{\text{synt}} \sim f(\tilde{Y}|X,Y) = \int f(\tilde{Y}|X,\theta) f(\theta|X,Y) \mathrm{d}\theta \propto \int f(\tilde{Y}|X,\theta) f(Y|X,\theta) f(\theta) \mathrm{d}\theta$$

   - Using the combining rule for correct uncertainty quantification
     - under both Bayesian and repeated sampling frameworks

     $\Rightarrow$ Some deep-learning approaches (e.g., table-GAN) overlook the uncertainty quantification

- Combining rule for uncertainty quantification (under repeated sampling)



$$\frac{\sqrt{n}(\hat{Q}-Q)}{\sqrt{U}} \to N(0,1)$$

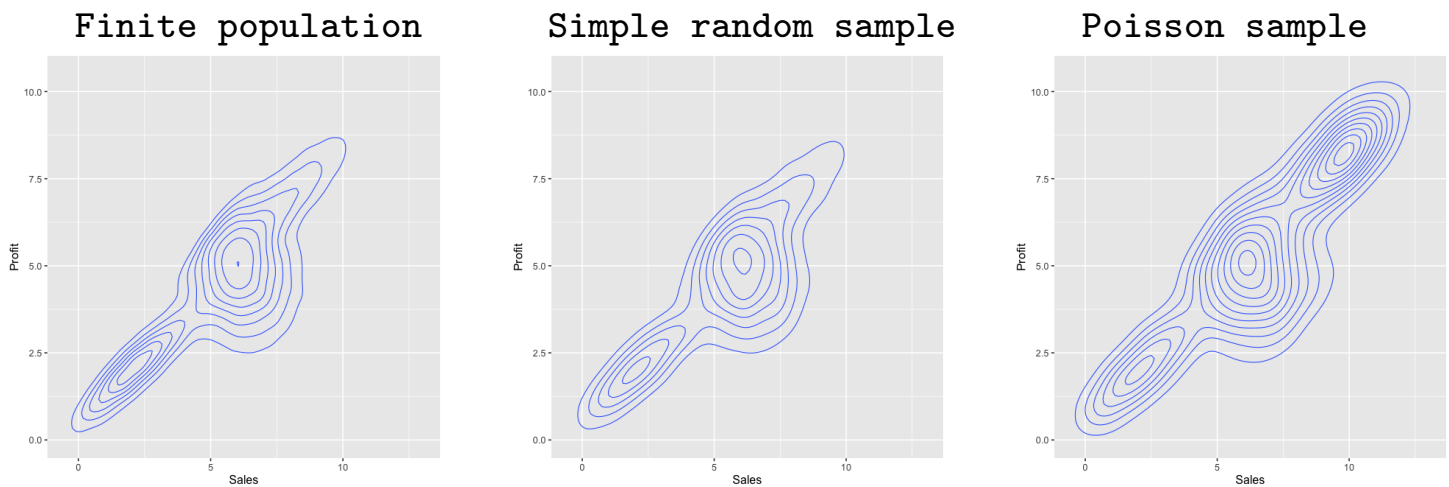
$$\frac{\sqrt{n}(\bar{Q}-Q)}{\sqrt{T}} \to N(0,1)$$

$\bar{Q}$ = Avg. of $\hat{Q}^{(l)}$

T = Avg. of $U^{(l)} + \frac{1}{m}\left(\text{Var. of } \hat{Q}^{(l)}\right)$

Synthesis for survey sampilng data

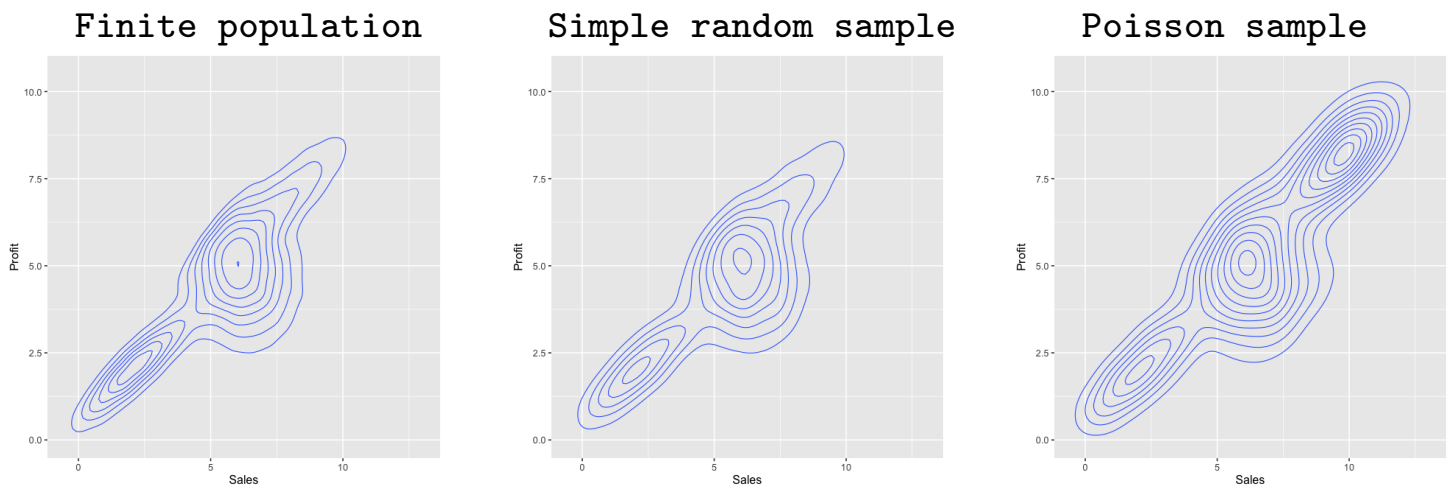
- The sampilng distribution of survey data often differs with the distributino of finite population.



- Some model-based approaches with survey weights
  1. Disregarding the survey weights, $\prod_{i=1}^{n} f(y_i|\theta) \ = \ ?$
  2. Reconstruct the finite pop. (bootstrap), $f(y_1, \ldots, y_N|\theta) = \prod_{i=1}^{N} f(\tilde{y}_i|\theta)$, where $\tilde{y}_i = y_i$ for sampled units and other $\tilde{y}_i$ are estimated/resampled.
  3. Using the pseudo likelihood, $f(y_1, \ldots, y_N|\theta) \ \approx \ \prod_{i=1}^{n} f(y_i|\theta)^{w_i}$.

Synthesis for survey sampilng data

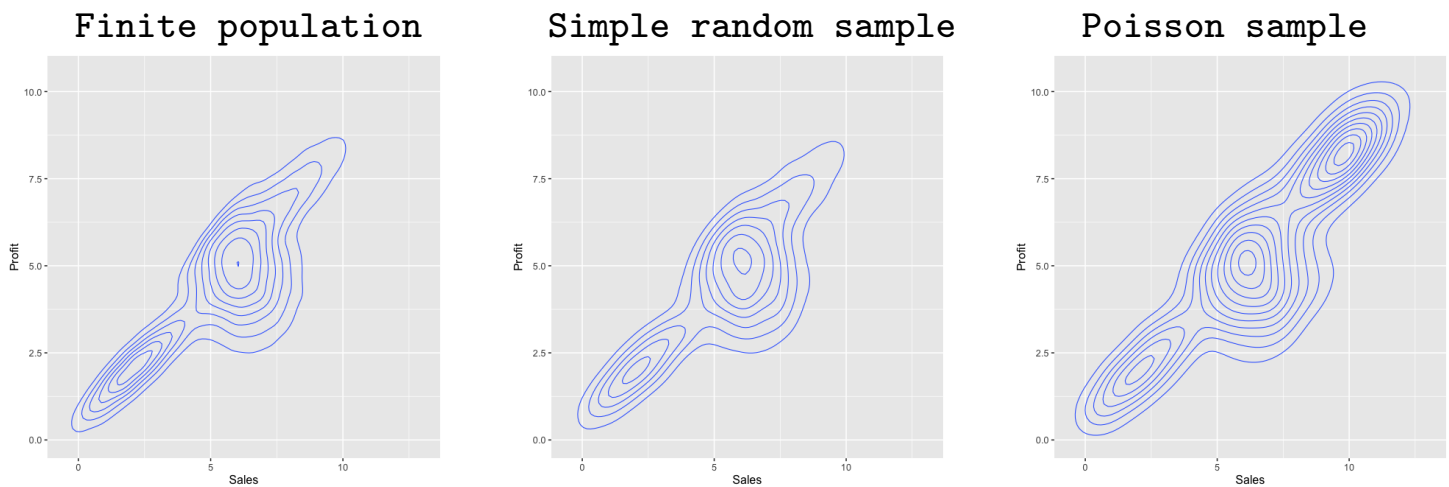
- The sampilng distribution of survey data often differs with the distributino of finite population.



- Some model-based approaches with survey weights
  1. Disregarding the survey weights, $\prod_{i=1}^{n} f(y_i|\theta) \ = \ ?$
  2. Reconstruct the finite pop. (bootstrap), $f(y_1, \ldots, y_N|\theta) = \prod_{i=1}^{N} f(\tilde{y}_i|\theta)$, where $\tilde{y}_i = y_i$ for sampled units and other $\tilde{y}_i$ are estimated/resampled.
  3. Using the pseudo likelihood, $f(y_1, \ldots, y_N|\theta) \ \approx \ \prod_{i=1}^{n} f(y_i|\theta)^{w_i}$.

Synthesis for survey sampilng data

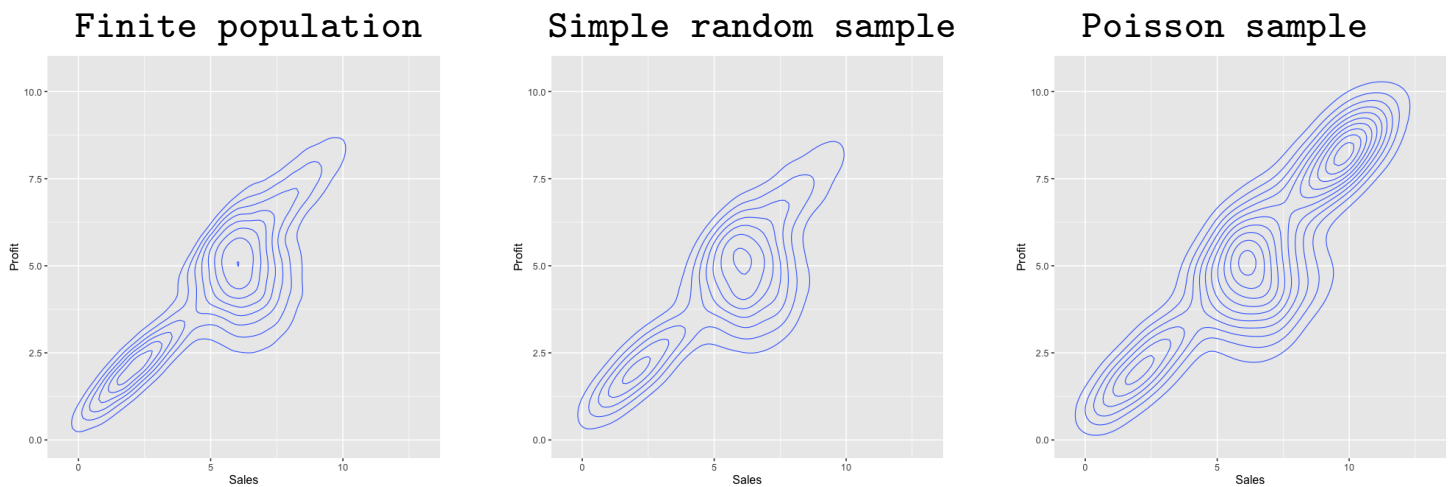
- The sampilng distribution of survey data often differs with the distributino of finite population.



- Some model-based approaches with survey weights
  1. Disregarding the survey weights, $\prod_{i=1}^{n} f(y_i|\theta) \ = \ ?$
  2. Reconstruct the finite pop. (bootstrap), $f(y_1, \ldots, y_N|\theta) = \prod_{i=1}^{N} f(\tilde{y}_i|\theta)$, where $\tilde{y}_i = y_i$ for sampled units and other $\tilde{y}_i$ are estimated/resampled.
  3. Using the pseudo likelihood, $f(y_1, \ldots, y_N|\theta) \ \approx \ \prod_{i=1}^{n} f(y_i|\theta)^{w_i}$.

Synthesis for survey sampilng data

- The sampilng distribution of survey data often differs with the distributino of finite population.



- Some model-based approaches with survey weights
  1. Disregarding the survey weights, $\prod_{i=1}^{n} f(y_i|\theta) \ = \ ?$
  2. Reconstruct the finite pop. (bootstrap), $f(y_1, \ldots, y_N|\theta) = \prod_{i=1}^{N} f(\tilde{y}_i|\theta)$, where $\tilde{y}_i = y_i$ for sampled units and other $\tilde{y}_i$ are estimated/resampled.
  3. Using the pseudo likelihood, $f(y_1, \ldots, y_N|\theta) \ \approx \ \prod_{i=1}^{n} f(y_i|\theta)^{w_i}$.

Bayesian pseudo posterior approach (Savitsky and Toth, 2016)

Assuming that $(w_i - 1)$ non-sampled units have the same values as a sampled unit y_i in evaluating the (pseudo) likelihood fn. $l^{\text{pse}}(\theta) = \prod_{i=1}^{n} f(y_i|\theta)^{w_i}$,

$$f^{\text{pse}}(\theta|\boldsymbol{y}_n, \boldsymbol{w}_n) = f^{\text{pse}}(\theta|y_1, \ldots, y_n, w_1, \ldots, w_n) \propto \prod_{i=1}^{n} f(y_i|\theta)^{w_i} \cdot f(\theta)$$

We proved that

1. $E(\theta|\text{Data})$ with f^{pse} is asymptotically unbiased. [Bernstein–Von Mises]

$$(n\boldsymbol{Q}_0)^{1/2} \; f^{\text{pse}}(\theta|\boldsymbol{y}) \rightarrow \mathcal{N}(\theta_0, \boldsymbol{I}) \text{ as } n \rightarrow \infty \text{ where } \boldsymbol{Q}_0 = -E_0\left[\nabla^2 l^{\text{pse}}(\theta)\right]$$

2. Posterior variance of θ is not close to the variance of the posterior mean for repeated sampling, i.e., $E(\hat{V}(\theta|\text{Data})) \neq V(\widehat{E}(\theta|\text{Data}))$ [Godambe information]

$$\left(n\boldsymbol{Q}_0\boldsymbol{P}_0^{-1}\boldsymbol{Q}_0\right)^{1/2}\left(\hat{\theta}_n^{\text{pse}} - \theta_0\right) \rightarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \text{ where } \boldsymbol{P}_0 = E_0\left[\nabla l^{\text{pse}}(\theta)\nabla l^{\top\text{pse}}(\theta)\right]$$

Suggestion: Variance-adjusted pseudo posterior

With the original pseudo posterior approach, $\boldsymbol{Q}_0 \neq \boldsymbol{P}_0$.

We suggest to use the power of the adjusted weights κw_i,

$$f^{\text{adj}}(\theta|y_1,\ldots,y_n,w_1,\ldots,w_n) \propto \prod_{i=1}^{n} f(y_i|\theta)^{\kappa w_i} \cdot f(\theta) \quad \text{where } \kappa = \frac{\sum_{j=1}^{n} w_j}{\sum_{j=1}^{n} w_j^2}.$$

1. With the adjusted weights, $\boldsymbol{Q}_0 = \boldsymbol{P}_0$, so the posterior mean with the adjusted pseudo likelihood follows

$$\sqrt{n}\left(\hat{\theta}_n^{\text{adj}} - \theta_0\right) \to \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right)$$

2. In SRS, the adjusted pseudo posterior becomes the posterior distribution disregarding the survey weights, i.e.,

$$\kappa w_i = \frac{\sum_{j=1}^{n} \frac{N}{n}}{\sum_{j=1}^{n} \frac{N^2}{n^2}} \frac{N}{n} = 1 \quad \Rightarrow \quad \prod_{i=1}^{n} f(y_i|\theta)^{\kappa w_i} \cdot f(\theta) = \prod_{i=1}^{n} f(y_i|\theta) \cdot f(\theta)$$

Simulation study result summary

Three synthetic data methods in comparison

1. **No weight**, ignoring survey weights, $\prod_{i=1}^{n} f(y_i|\theta) \cdot f(\theta)$.
2. **Pseudo** posterior with the original survey weights, $\prod_{i=1}^{n} f(y_i|\theta)^{w_i} \cdot f(\theta)$.
3. **Adjusted** pseudo posterior, $\prod_{i=1}^{n} f(y_i|\theta)^{\kappa w_i} \cdot f(\theta)$.

Sampling Methods		**No weight**	**Pseudo**	**Adjusted**
Simple Random Sampling	$E(\hat{\bar{Y}}_1) - \bar{Y}_1$	0.00	0.00	0.00
	$V(\hat{\bar{Y}}_1)$	0.027	0.027	0.028
	$E(\hat{V}(\hat{\bar{Y}}_1))$	0.025	0.001	0.025
	$95\%\ C.I\ coverage$	0.928	0.286	0.922
Poisson Sampling	$E(\hat{\bar{Y}}_1) - \bar{Y}_1$	2.02	0.00	0.00
	$V(\hat{\bar{Y}}_1)$	0.030	0.031	0.031
	$E(\hat{V}(\hat{\bar{Y}}_1))$	0.025	0.001	0.027
	$95\%\ C.I\ coverage$	0.000	0.298	0.924

Concluding remarks

1. Disregarding sampling weights results in biased estimation when the sample is collected with unequal probability sampling.
2. The (original) pseudo posterior approach results in variance underestimation.
3. The suggested pseudo likelihood approach with **the adjusted weight** results in correct estimation with imputed (and synthetic) data.

- Note: William and Savitsky (2021) suggested a post-processing **after** MCMC.
  In case of needing imputation, the adjustment needs to be given **during** MCMC as suggested here.

Thank you!

Contact Information

Hang Kim (hang.kim@uc.edu)

Division of Statistics and Data Science
Department of Mathematical Sciences
University of Cincinnati

University of CINCINNATI