Variance adjusted inference for unequal probability sample with application to imputation and data synthesis

Hang J. Kim

Division of Statistics and Data Science

University of Cincinnati

presented at 2024 Privacy and Public Policy Conference

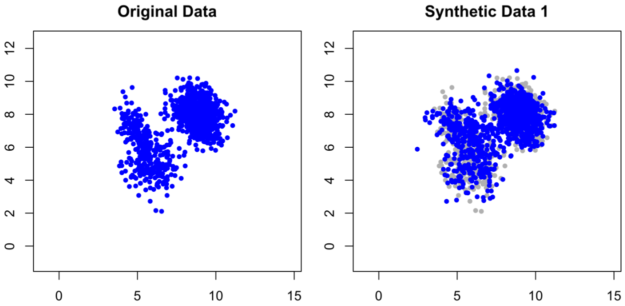
Georgetown University, Washington, DC

September 14, 2024

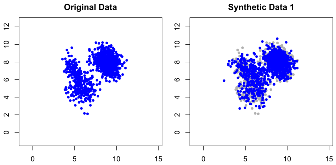
Synthetic data generation with a probabilistic model

- Synthetic data: proposed by Rubin (1993) assuming probabilistic models
  - Current, the term is used in broader sense

1. Assume (a family of) the distribution of the original data: $f(y_{\text{orig}}|\theta)$
2. Learn the distribution of the original data: $\hat{\theta}$ or $f(\theta|y_{\text{orig}})$
3. Randomly generate synthetic values: $f(\tilde{y}_{\text{synt}}|y_{\text{orig}}) = \int f(\tilde{y}_{\text{synt}}|\theta)f(\theta|y_{\text{orig}})\mathrm{d}\theta$



Why is variance estimation with synthetic data important?



- Jerry Reiter (Duke) and colleagues have showed synthetic data generated with nonparametric Bayesian models support well user's various analyses:
  - plausible point estimators, e.g., regression coefficients $\hat{\beta}$
  - and honest **variance estimator**, e.g., $\widehat{V}(\hat{\beta})$

    $V(\hat{\theta}_{\text{synt}}) = V(\hat{\theta}_{\text{orig}}) + U$ where U is uncertinty due to synthesis
- Some data privacy methods cannot measure U or provide incorrect $V(\hat{\theta}_{\text{synt}})$
  - Hypotheses testing results in false positive

    $\Rightarrow$ reproducibility issues in scientific research

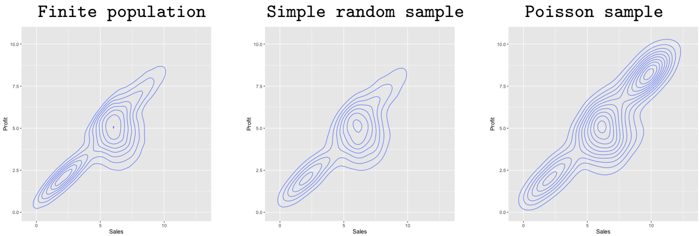
University of CINCINNATI

ADD HYPOTHESIS TESTING

Modeling survey sampilng data

- Unequal probability sampling

  : Distribution of survey sample often differs from that of finite population.

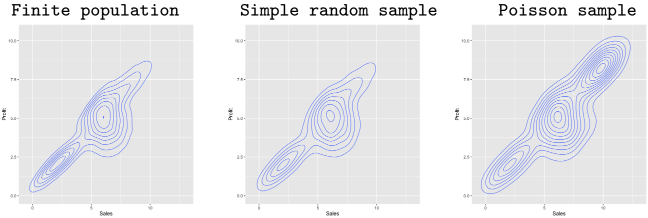


  - e.g., establishment surveys: Large companies receive high inclusion probability

    $\Rightarrow$ The variance of total sales gets lower.

- Survey weights w_i are used to derive a correct (design-unbiased) estimtor.
- Assume that an agency wants to generate synthetic (finite) populations

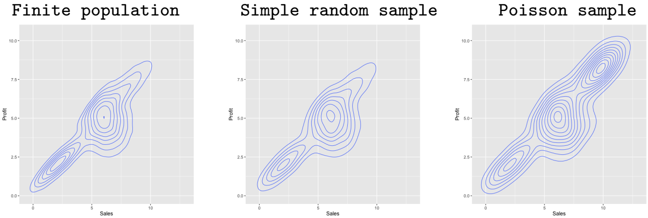
What likelihood functions need to be used?



Some (probabilistic) model-based approaches with survey weights

1. Disregarding the survey weights, $\prod_{i=1}^{n} f(y_i|\theta) \ = ?$

What likelihood functions need to be used?

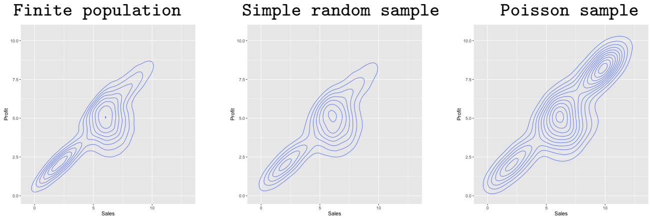


Some (probabilistic) model-based approaches with survey weights

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.

What likelihood functions need to be used?



Some (probabilistic) 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. $l^{\text{pse}}(\theta) = \prod_{i=1}^{n} f(y_i|\theta)^{w_i}$,

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

- The pseudo posterior approach generates synthetic data that result in
  - consistent point estimator $\hat{\theta}$ but
  - underestimated variance estimator $E[\hat{V}(\hat{\theta})] < V(\hat{\theta})$.
- Solutions
  - William and Savitsky (2021) suggested a post-processing **after** MCMC.
  - We propose an adjustment given **during** MCMC, so
    - correct synthetic populations are generated during MCMC, and
    - handle imcomplete survey data with missing records.

For the pseudo posterior distribution

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

we proved that

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

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

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]

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

* Sandwich estimator for the misspecified likelihood

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

Suggestion: Variance-adjusted pseudo posterior

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

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

Then,

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

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

2. With the adjusted weights, $\boldsymbol{P}_0 = \boldsymbol{Q}_0 = -E_0\left[\nabla^2 l^{\text{adj}}(\theta)\right]$, so the posterior mean with the adjusted pseudo likelihood follows

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

Suggestion: Variance-adjusted pseudo posterior

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

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

Then,

3. 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: Comparison three synthesis methods

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	$E(\hat{\hat{Y}}_1) - \bar{Y}_1$	0.00	0.00	0.00
	$V(\hat{\hat{Y}}_1)$	0.027	0.027	0.028
	$E(\hat{V}(\hat{\hat{Y}}_1))$	0.025	0.001	0.025
	$95\%\ C.I\ coverage$	0.928	0.286	0.922
Poisson Sampling	$E(\hat{\hat{Y}}_1) - \bar{Y}_1$	2.02	0.00	0.00
	$V(\hat{\hat{Y}}_1)$	0.030	0.031	0.031
	$E(\hat{V}(\hat{\hat{Y}}_1))$	0.025	0.001	0.027
	$95\%\ C.I\ coverage$	0.000	0.298	0.924

Simulation study: Comparison three synthesis methods

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	$E(\hat{\hat{Y}}_1) - \bar{Y}_1$	0.00	0.00	0.00
	$V(\hat{\hat{Y}}_1)$	0.027	0.027	0.028
	$E(\hat{V}(\hat{\hat{Y}}_1))$	0.025	0.001	0.025
	$95\%\ C.I\ coverage$	0.928	0.286	0.922
Poisson Sampling	$E(\hat{\hat{Y}}_1) - \bar{Y}_1$	2.02	0.00	0.00
	$V(\hat{\hat{Y}}_1)$	0.030	0.031	0.031
	$E(\hat{V}(\hat{\hat{Y}}_1))$	0.025	0.001	0.027
	$95\%\ C.I\ coverage$	0.000	0.298	0.924

Simulation study: Comparison three synthesis methods

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	$E(\hat{\hat{Y}}_1) - \bar{Y}_1$	0.00	0.00	0.00
	$V(\hat{\hat{Y}}_1)$	0.027	0.027	0.028
	$E(\hat{V}(\hat{\hat{Y}}_1))$	0.025	0.001	0.025
	$95\%\ C.I\ coverage$	0.928	0.286	0.922
Poisson Sampling	$E(\hat{\hat{Y}}_1) - \bar{Y}_1$	2.02	0.00	0.00
	$V(\hat{\hat{Y}}_1)$	0.030	0.031	0.031
	$E(\hat{V}(\hat{\hat{Y}}_1))$	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.

Thank you!

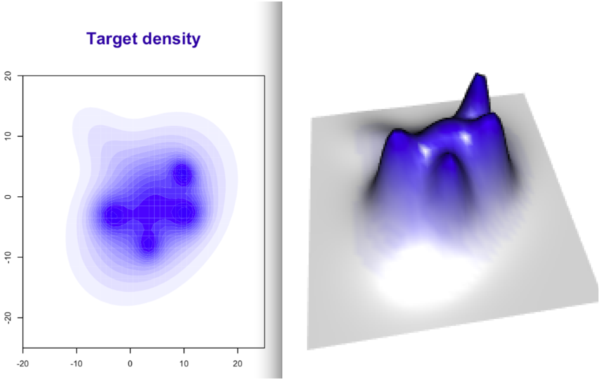
Contact Information

Hang Kim (hang.kim@uc.edu)

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

Appendix: Development in joint modeling

- What distribution is good to fit the empirical density?

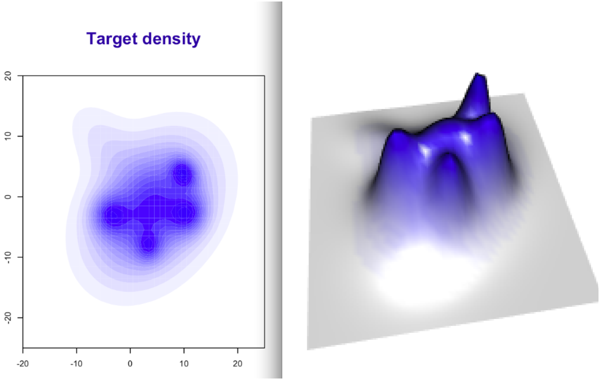


$\Rightarrow$

-
-

Appendix: Development in joint modeling

- What distribution is good to fit the empirical density?

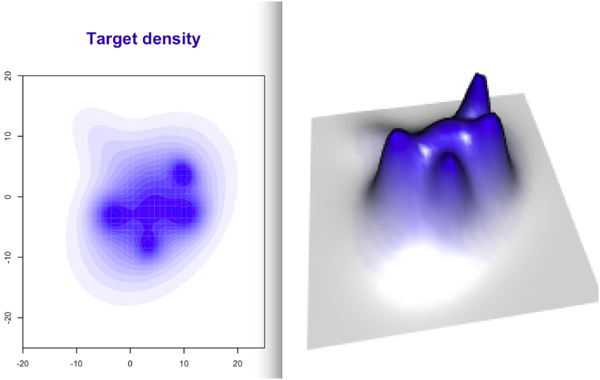


$\Rightarrow$ Mixture distribution $f(y_i|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{6} w_k \ \mathsf{N}\left(y_i; \mu_k, \Sigma_k\right)$

-
-

Appendix: Development in joint modeling

- What distribution is good to fit the empirical density?

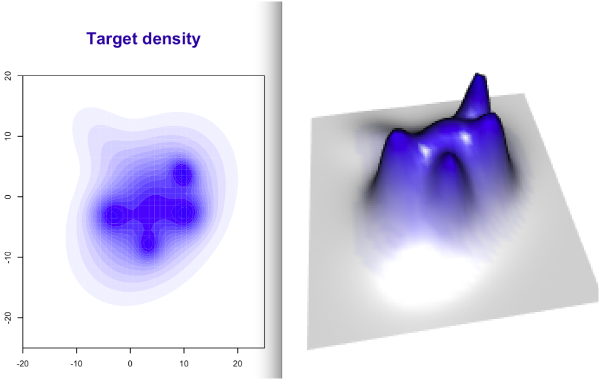


$\Rightarrow$ Mixture distribution $f(y_i|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{6} w_k \; \mathrm{N}\left(y_i; \mu_k, \Sigma_k\right)$

- Estimated by a nonparameteric Bayesian model
-

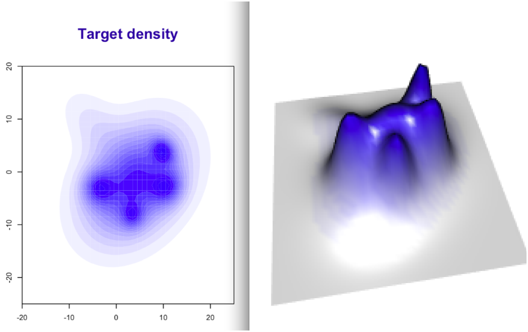
Appendix: Development in joint modeling

- What distribution is good to fit the empirical density?



$\Rightarrow$ Mixture distribution $f(y_i|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{6} w_k \; \mathsf{N}\left(y_i; \mu_k, \Sigma_k\right)$

- Estimated by a nonparameteric Bayesian model
- Jerry Reiter (Duke) `http://www2.stat.duke.edu/~jerry/papers.html`



⇒ Generated from a mixture of **6** multivariate normal distributions [1]

$$f(y_i|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{6} w_k \ \mathsf{N}\left(y_i; \mu_k, \Sigma_k\right)$$

⇒ Also represented by using a membership indicator $z_i \in \{1, \ldots, 6\}$

$$f(z_i|\boldsymbol{w}) \sim \text{Categorical}(w_1, \ldots, w_6), \qquad f(y_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, z_i) \sim N(y_i; \mu_{z_i}, \Sigma_{z_i})$$

such that $f(y_i|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int f(z_i|\boldsymbol{w}) f(y_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, z_i) \mathrm{d} z_i$

Nonparametric Bayes: Dirichlet process Gaussian mixture

- Challenges for a mixture of normal (Gaussian) distributions
  1. Simultaneous estimation w_k, μ_k, Σ_k for $k = 1, \ldots, K$
  2. Effective number of mixture components (how many normal kernels?)

$\Rightarrow$ Dirichlet process: Let **data** inform the decision

- Dirichlet process (DP) prior: Stick-breaking representation

$$
\begin{aligned}
w_k &= \nu_k \prod_{g<k} (1 - \nu_g) \quad \text{for } k = 1, \ldots, K \\
\nu_k | \alpha &\sim \text{Beta}(1, \alpha) \quad \text{for } k = 1, \ldots, K-1; \;\; \nu_K = 1, \\
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha).
\end{aligned}
$$

The DP Gaussian mixture is a famous form of nonparametric Bayesian models.

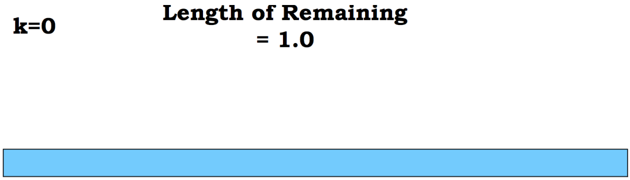
Stick-breaking Representation (Sethuraman 1994)

- Automatically determines w_k, reflecting information from $\boldsymbol{x}_i$

$$p_k \sim \text{Beta}(1, \alpha)$$

$$w_1 = p_1, \quad w_2 = p_1\,(1 - w_1), \quad w_3 = p_2\,(1 - w_1 - w_2), \quad \ldots$$

$$\left(1 - \sum_{g=1}^{k-1} w_g\right) \qquad\qquad p_k$$



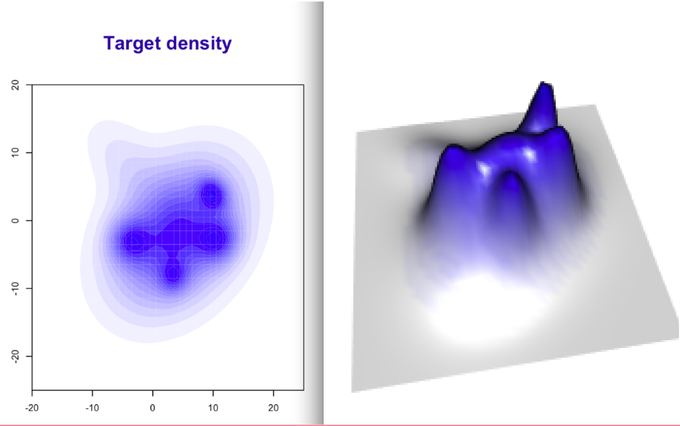
w_k

- Concentration parameter α
  - Larger $\alpha \to$ smaller $E(p_k)$ ... $\to$ more number of components used

DP mixture model decides

1. how many components are to be used
2. contribution of each component to explain the empirical dist'n
3. location and shape of each normal component

based on data information



Nonparametric Bayesian Data Synthesis for Cont. Data

1. Likelihood: Mixture Normals

$$p(\boldsymbol{y}_i|\mathcal{A}) \propto \left(\sum_{k=1}^{K} w_k N(\boldsymbol{y}_i|\mu_k, \Sigma_k)\right) I(\boldsymbol{y}_i \in \mathcal{A})$$

   - $\mathcal{A}$: support of original values (예: 남자 종사자수 $\leq$ 총 종사자수)

2. Prior for w_k: Dirichlet process (DP) model
   - $w_1 = p_1$
   - $w_k = p_k \left(1 - \sum_{g=1}^{k-1} w_g\right)$ for $k = 2, \ldots, K$
   - $p_k \sim$ Beta$(1, \alpha)$

3. Conjugate priors for μ_k and Σ_k: Normal-Inverse-Wishart

4. Weak priors for other hyperparameters

MCMC Steps

Most updates are based on Gibbs, i.e., closed forms of conditional distributions.

1. Update* $\{\mu_k, \Sigma_k\}$ given $Y_n = \{\boldsymbol{y}_i; \boldsymbol{y}_i \in \mathcal{A}\}$ and $Z_n = \{z_1, \ldots, z_n\}$.
2. Update the membership indicator z_i
3. Update component weight $\boldsymbol{w} = (w_1, \ldots, w_K)$
4. Generate synthetic data $\tilde{y}$ given $\{\mu_k, \Sigma_k, w_k\}$
5. Repeat Step 1 – 4