# Fractional hot deck imputation for multivariate missing data in survey sampling

Jae Kwang Kim     Wayne A. Fuller

Iowa State University

October 17, 2014

# Fractional imputation

Features

- Split the record with missing item into $m(> 1)$ imputed values
- Assign fractional weights
- The final product is a single data file with size $\leq nm$.
- For variance estimation, the fractional weights are replicated.

Example ($n = 10$)

| ID | Weight | $y_1$ | $y_2$ |
|---|---|---|---|
| 1 | $w_1$ | $y_{1,1}$ | $y_{1,2}$ |
| 2 | $w_2$ | $y_{2,1}$ | M |
| 3 | $w_3$ | M | $y_{3,2}$ |
| 4 | $w_4$ | $y_{4,1}$ | $y_{4,2}$ |
| 5 | $w_5$ | $y_{5,1}$ | $y_{5,2}$ |
| 6 | $w_6$ | $y_{6,1}$ | $y_{6,2}$ |
| 7 | $w_7$ | M | $y_{7,2}$ |
| 8 | $w_8$ | M | M |
| 9 | $w_9$ | $y_{9,1}$ | $y_{9,2}$ |
| 10 | $w_{10}$ | $y_{10,2}$ | $y_{10,2}$ |

M: Missing

# Fractional imputation

**Fractional Imputation Idea**

If both $y_1$ and $y_2$ are categorical, then fractional imputation is easy to apply.

- We have only finite number of possible values.
- Imputed values = possible values
- The fractional weights are the conditional probabilities of the possible values given the observations.
- Can use "EM by weighting" method of Ibrahim (1990) to compute the fractional weights.

# Fractional imputation

Example ($y_1, y_2$: dichotomous, taking 0 or 1)

| ID | Weight | $y_1$ | $y_2$ |
|---|---|---|---|
| 1 | $w_1$ | $y_{1,1}$ | $y_{1,2}$ |
| 2 | $w_2 w_{2,1}^*$ | $y_{2,1}$ | 0 |
| | $w_2 w_{2,2}^*$ | $y_{2,1}$ | 1 |
| 3 | $w_3 w_{3,1}^*$ | 0 | $y_{3,2}$ |
| | $w_3 w_{3,2}^*$ | 1 | $y_{3,2}$ |
| 4 | $w_4$ | $y_{4,1}$ | $y_{4,2}$ |
| 5 | $w_5$ | $y_{5,1}$ | $y_{5,2}$ |

# Fractional imputation

Example ($y_1, y_2$: dichotomous, taking 0 or 1)

| ID | Weight | $y_1$ | $y_2$ |
|---|---|---|---|
| 6 | $w_6$ | $y_{6,1}$ | $y_{6,2}$ |
| 7 | $w_7 w_{7,1}^*$ | 0 | $y_{7,2}$ |
| | $w_7 w_{7,2}^*$ | 1 | $y_{7,2}$ |
| 8 | $w_8 w_{8,1}^*$ | 0 | 0 |
| | $w_8 w_{8,2}^*$ | 0 | 1 |
| | $w_8 w_{8,3}^*$ | 1 | 0 |
| | $w_8 w_{8,4}^*$ | 1 | 1 |
| 9 | $w_9$ | $y_{9,1}$ | $y_{9,2}$ |
| 10 | $w_{10}$ | $y_{10,1}$ | $y_{10,2}$ |

Example (Cont'd)

- E-step: Fractional weights are the conditional probabilities of the imputed values given the observations.

$$
\begin{aligned}
w_{ij}^{*} &= \hat{P}(y_{i,mis}^{*(j)} \mid y_{i,obs}) \\
&= \frac{\hat{\pi}(y_{i,obs}, y_{i,mis}^{*(j)})}{\sum_{l=1}^{M_i} \hat{\pi}(y_{i,obs}, y_{i,mis}^{*(l)})}
\end{aligned}
$$

where $(y_{i,obs}, y_{i,mis})$ is the (observed, missing) part of $y_i = (y_{i1}, \cdots, y_{i,p})$.

- M-step: Update the joint probability using the fractional weights.

$$
\hat{\pi}_{ab} = \frac{1}{\hat{N}} \sum_{i=1}^{n} \sum_{j=1}^{M_i} w_i w_{ij}^{*} I(y_{i,1}^{*(j)} = a, y_{i,2}^{*(j)} = b)
$$

with $\hat{N} = \sum_{i=1}^{n} w_i$.

# Fractional imputation

Example (Cont'd) Variance estimation

- Recompute the fractional weights for each replication
- Apply the same EM algorithm using the replicated weights.
  - E-step: Fractional weights are the conditional probabilities of the imputed values given the observations.

$$w_{ij}^{*(k)} = \frac{\hat{\pi}^{(k)}(y_{i,obs}, y_{i,mis}^{*(j)})}{\sum_{l=1}^{M_i} \hat{\pi}^{(k)}(y_{i,obs}, y_{i,mis}^{*(l)})}$$

  - M-step: Update the joint probability using the fractional weights.

$$\hat{\pi}_{ab}^{(k)} = \frac{1}{\hat{N}^{(k)}} \sum_{i=1}^{n} \sum_{j=1}^{M_i} w_i^{(k)} w_{ij}^{*(k)} I(y_{i,1}^{*(j)} = a, y_{i,2}^{*(j)} = b)$$

where $\hat{N}^{(k)} = \sum_{i=1}^{n} w_i^{(k)}$.

# Fractional imputation

Example (Cont'd) Final Product

| | | | | Replication Weights | | | |
|---|---|---|---|---|---|---|---|
| Weight | $x$ | $y_1$ | $y_2$ | Rep 1 | Rep 2 | $\cdots$ | Rep $L$ |
| $w_1$ | $x_1$ | $y_{1,1}$ | $y_{1,2}$ | $w_1^{(1)}$ | $w_1^{(2)}$ | $\cdots$ | $w_1^{(L)}$ |
| $w_2 w_{2,1}^*$ | $x_2$ | $y_{2,2}$ | 0 | $w_2^{(1)} w_{2,1}^{*(1)}$ | $w_2^{(2)} w_{2,1}^{*(2)}$ | $\cdots$ | $w_2^{(L)} w_{2,1}^{*(L)}$ |
| $w_2 w_{2,2}^*$ | $x_2$ | $y_{2,2}$ | 1 | $w_2^{(1)} w_{2,2}^{*(1)}$ | $w_2^{(2)} w_{2,1}^{*(2)}$ | $\cdots$ | $w_2^{(L)} w_{2,2}^{*(L)}$ |
| $w_3 w_{3,1}^*$ | $x_3$ | 0 | $y_{3,2}$ | $w_3^{(1)} w_{3,1}^{*(1)}$ | $w_3^{(2)} w_{3,1}^{*(2)}$ | $\cdots$ | $w_3^{(L)} w_{3,1}^{*(L)}$ |
| $w_3 w_{3,2}^*$ | $x_3$ | 1 | $y_{3,2}$ | $w_3^{(1)} w_{3,2}^{*(1)}$ | $w_3^{(2)} w_{3,2}^{*(2)}$ | $\cdots$ | $w_3^{(L)} w_{3,2}^{*(L)}$ |
| $w_4$ | $x_4$ | $y_{4,1}$ | $y_{4,2}$ | $w_4^{(1)}$ | $w_4^{(2)}$ | $\cdots$ | $w_4^{(L)}$ |
| $w_5$ | $x_5$ | $y_{5,1}$ | $y_{5,2}$ | $w_5^{(1)}$ | $w_5^{(2)}$ | $\cdots$ | $w_5^{(L)}$ |
| $w_6$ | $x_6$ | $y_{6,1}$ | $y_{6,2}$ | $w_6^{(1)}$ | $w_6^{(2)}$ | $\cdots$ | $w_6^{(L)}$ |

# Fractional imputation

Example (Cont'd) Final Product

| | | | | Replication Weights | | | |
|---|---|---|---|---|---|---|---|
| Weight | $x$ | $y_1$ | $y_2$ | Rep 1 | Rep 2 | | Rep L |
| $w_7 w_{7,1}^*$ | $x_7$ | 0 | $y_{7,2}$ | $w_7^{(1)} w_{7,1}^{*(1)}$ | $w_7^{(2)} w_{7,1}^{*(2)}$ | | $w_7^{(L)} w_{7,1}^{*(L)}$ |
| $w_7 w_{7,2}^*$ | $x_7$ | 1 | $y_{7,2}$ | $w_7^{(1)} w_{7,2}^{*(1)}$ | $w_7^{(2)} w_{7,2}^{*(2)}$ | | $w_7^{(L)} w_{7,2}^{*(L)}$ |
| $w_8 w_{8,1}^*$ | $x_8$ | 0 | 0 | $w_8^{(1)} w_{8,1}^{*(1)}$ | $w_8^{(1)} w_{8,1}^{*(2)}$ | | $w_8^{(L)} w_{8,1}^{*(L)}$ |
| $w_8 w_{8,2}^*$ | $x_8$ | 0 | 1 | $w_8^{(1)} w_{8,2}^{*(1)}$ | $w_8^{(2)} w_{8,2}^{*(2)}$ | | $w_8^{(L)} w_{8,2}^{*(L)}$ |
| $w_8 w_{8,3}^*$ | $x_8$ | 1 | 0 | $w_8^{(1)} w_{8,3}^{*(1)}$ | $w_8^{(2)} w_{8,3}^{*(2)}$ | | $w_8^{(L)} w_{8,3}^{*(L)}$ |
| $w_8 w_{8,4}^*$ | $x_8$ | 1 | 1 | $w_8^{(1)} w_{8,4}^{*(1)}$ | $w_8^{(1)} w_{8,4}^{*(2)}$ | | $w_8^{(L)} w_{8,4}^{*(L)}$ |
| $w_9$ | $x_9$ | $y_{9,1}$ | $y_{9,2}$ | $w_9^{(1)}$ | $w_9^{(2)}$ | $\cdots$ | $w_9^{(L)}$ |
| $w_{10}$ | $x_{10}$ | $y_{10,1}$ | $y_{10,2}$ | $w_{10}^{(1)}$ | $w_{10}^{(2)}$ | $\cdots$ | $w_{10}^{(L)}$ |

# Fractional hot deck imputation

Goals

- Fractional hot deck imputation of size $m$. The final product is a single data file with size $\leq n \cdot m$.
- Preserves correlation structure
- Variance estimation relatively easy
- Can handle domain estimation (but we do not know which domains will be used.)

# Fractional hot deck imputation

Parametric model approach

- In the parametric model approach, imputed values can be generated from $f(y_{i,mis} \mid y_{i,obs}, \hat{\theta})$ where $\hat{\theta}$ is the MLE of $\theta$.
- Kim (2011) proposed parametric fractional imputation (PFI) which is based on importance sampling idea

$$w_{ij}^* \propto \frac{f(y_{i,obs}, y_{i,mis}^{*(j)}; \hat{\theta})}{h(y_{i,mis}^{*(j)} \mid y_{i,obs})}$$

with $\sum_{j=1}^{m} w_{ij}^* = 1$, where $y_{i,mis}^{*(1)}, \cdots, y_{i,mis}^{*(m)} \sim h(y_{i,mis} \mid y_{i,obs})$.

# Fractional hot deck imputation

Idea

- In hot deck imputation, we can make a nonparametric approximation of $f(\cdot)$ using a finite mixture model

$$f(y_{i,mis} \mid y_{i,obs}) = \sum_{g=1}^{G} \pi_g(y_{i,obs}) f_g(y_{i,mis}), \tag{1}$$

where $\pi_g(y_{i,obs}) = P(z_i = g \mid y_{i,obs})$, $f_g(y_{i,mis}) = f(y_{i,mis} \mid z = g)$ and $z$ is the latent variable associated with imputation cell.

- To satisfy the above approximation, we need to find $z$ such that

$$f(y_{i,mis} \mid z_i, y_{i,obs}) = f(y_{i,mis} \mid z_i).$$

# Fractional hot deck imputation

## Imputation cell

- Assume $p$-dimensional survey items: $Y = (Y_1, \cdots, Y_p)$
- For each item $k$, create a transformation of $Y_k$ into $Z_k$, a discrete version of $Y_k$ based on the sample quantiles among respondents.
- If $y_{i,k}$ is missing, then $z_{i,k}$ is also missing.
- Imputation cells are created based on the observed value of $z_i = (z_{i,1}, \cdots, z_{i,p})$.
- Expression (1) can be written as

$$f(y_{i,mis} \mid y_{i,obs}) = \sum_{z_{mis}} P(z_{i,mis} = z_{mis} \mid z_{i,obs}) f(y_{i,mis} \mid z_{mis}), \qquad (2)$$

where $z_i = (z_{i,obs}, z_{i,mis})$ similarly to $y_i = (y_{i,obs}, y_{i,mis})$.

# Fractional hot deck imputation

Estimation of cell probability

- Let $\{z_1, \cdots, z_G\}$ be the support $z$, which is the same as the sample support of $z$ from the full respondents.
- Cell probability $\pi_g = P(z = z_g)$.
- For each unit $i$, we only observe $z_{i,obs}$.
- Use EM algorithm for categorical missing data to estimate $\pi_g$.

# Fractional hot deck imputation

Two-stage sampling for fractional hot deck imputation of size $m$

- Stage 1: Given $z_{i,obs}$, imputed $m$ values of $z_{i,mis}$, denoted by $z_{i,mis}^{*(1)}, \cdots, z_{i,mis}^{*(m)}$, from the estimated conditional probability $P(z_{i,mis} \mid z_{i,obs})$.
- Stage 2: For each imputed cell $z_i^{*(j)} = (z_{i,obs}, z_{i,mis}^{*(j)})$, the imputed value for $y_{i,mis}$ is randomly chosen among the full respondents in the same cell. (Joint hot deck within imputation cell)

# Fractional hot deck imputation

Variance estimation

- Kim and Fuller (2004) proposed a variance estimation method for fractional imputation under the cell mean model.
- Kim, Fuller, and Bell (2011) applied the method to variance estimation for income estimates in the 2000 Census long form data.

# Example : Jackknife for Fractional Imputation in Kim and Fuller (2004)

| Unit | $w_i w_{ij}^*$ | $Y_1$ | $Y_2$ | $w_i^{(1)} w_{ij}^{*(1)}$ | $w_i^{(2)} w_{ij}^{*(2)}$ | $\cdots$ | $w_i^{(5)} w_{ij}^{*(5)}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.10 | 2 | 1 | 0 | 0.111 | | 0.111 |
| 2 | 0.10 | 2 | 2 | 0.111 | 0 | | 0.111 |
| 3 | 0.10 | 1 | 3 | 0.111 | 0.111 | | 0.111 |
| 4 | 0.10 | 4 | 4 | 0.111 | 0.111 | | 0.111 |
| 5 | 0.10 | 2 | 5 | 0.111 | 0.111 | | 0 |
| 6 | 0.05 | 3 | 1* | $0.111\,(0.5-\phi_1)$ | 0.055 | | $0.111\,(0.5+\phi_5)$ |
| | 0.05 | 3 | 5* | $0.111\,(0.5+\phi_1)$ | 0.055 | | $0.111\,(0.5-\phi_5)$ |
| 7 | 0.10 | 5 | 3 | 0.111 | 0.111 | | 0.111 |
| 8 | 0.10 | 7 | 6 | 0.111 | 0.111 | | 0.111 |
| 9 | 0.10 | 8 | 9 | 0.111 | 0.111 | | 0.111 |
| 10 | 0.05 | 5* | 3* | 0.055 | 0.055 | | 0.055 |
| | 0.05 | 7* | 6* | 0.056 | 0.056 | | 0.056 |

# Example (Continued)

| Unit | $w_i w_{ij}^*$ | $Y_1$ | $Y_2$ | $w_i^{(6)} w_{ij}^{*(6)}$ | $w_i^{(7)} w_{ij}^{*(7)}$ | $\cdots$ | $w_i^{(10)} w_{ij}^{*(10)}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.10 | 2 | 1 | 0.111 | 0.111 | | 0.111 |
| 2 | 0.10 | 2 | 2 | 0.111 | 0.111 | | 0.111 |
| 3 | 0.10 | 1 | 3 | 0.111 | 0.111 | | 0.111 |
| 4 | 0.10 | 4 | 4 | 0.111 | 0.111 | | 0.111 |
| 5 | 0.10 | 2 | 5 | 0.111 | 0.111 | | 0.111 |
| 6 | 0.05 | 3 | 1* | 0 | 0.055 | | 0.055 |
| | 0.05 | 3 | 5* | 0 | 0.056 | | 0.056 |
| 7 | 0.10 | 5 | 3 | 0.111 | 0 | | 0.111 |
| 8 | 0.10 | 7 | 6 | 0.111 | 0.111 | | 0.111 |
| 9 | 0.10 | 8 | 9 | 0.111 | 0.111 | | 0.111 |
| 10 | 0.05 | 5* | 3* | 0.55 | $0.111\,(0.5-\phi_7)$ | | 0 |
| | 0.05 | 7* | 6* | 0.56 | $0.111\,(0.5+\phi_7)$ | | 0 |

# Variance estimation for fractional imputation

- Variance estimator is a function of $\phi_k$'s :

$$\hat{V}_\phi = \sum_{k=1}^{L} c_k \left( \sum_{i \in A_R} \alpha_{\phi,i}^{(k)} y_i - \sum_{i \in A_R} \alpha_i y_i \right)^2$$

- Naive variance estimator ( $\phi_k \equiv 0$ ) : Underestimation
- Increasing the $\phi_k$ will increase the value of variance estimator
- How to decide $\phi_k$ ?

$$E\left(\hat{V}_\phi\right) - Var\left(\hat{\theta}_I\right) = E\left\{ \sum_{g=1}^{G} \sum_{i \in A_R} \left[ \sum_{k=1}^{L} c_k \left( \alpha_{\phi,i}^{(k)} - \alpha_i \right)^2 - \alpha_i^2 \right] \sigma_g^2 \right\}$$

# Variance estimation for fractional imputation

- Kim and Fuller (2004) showed that if

$$\sum_{i \in A_R} w_{ij}^{*(k)} = 1 \qquad (C.1)$$

and

$$\sum_{i \in A_{Rg}} \sum_{k=1}^{L} c_k \left( \alpha_i^{(k)} - \alpha_i \right)^2 = \sum_{i \in A_{Rg}} \alpha_i^2, \qquad (C.2)$$

then the replication variance estimator defined by

$$\hat{V}_I = \sum_{k=1}^{L} c_k \left( \hat{\theta}_I^{(k)} - \hat{\theta}_I \right)^2,$$

where $\hat{\theta}_I^{(k)} = \sum_{i \in A_R} \alpha_i^{(k)} y_i$, is unbiased for the total variance under the cell mean model.

- Three variables of size $n = 300$ are generated:

$$\begin{aligned} Y_1 &\sim U(0,2) \\ Y_2 &= 1 + Y_1 + e_2 \\ Y_3 &= 2 + Y_1 + 0.5y_2 + 0.5e_3 \end{aligned}$$

  where $e_2$ and $e_3$ are independently generated from a standard normal distribution truncated outside $[-3, 3]$.

- Response indicator functions for $Y_1$, $Y_2$, $Y_3$:

$$\delta_1, \delta_2, \delta_3 \stackrel{i.i.d.}{\sim} \text{Bernoulli}(0.7)$$

- Interested in $\theta_1 = E(Y_1)$, $\theta_2 = E(Y_2)$, $\theta_3 = E(Y_3)$, $\theta_4 = P(Y_1 < 1, Y_2 < 2)$, and $\theta_5 = E(Y_3 \mid D = 1)$ with $D \sim Bernoulli(0.3)$.

# Simulation study

## Simulation Setup

- $B = 2,000$ simulation samples.
- Multivariate fractional hot deck imputation was used with $m = 5$ fractional imputation.
- Categorical transformation (with 4 categories) was used to each of $Y_1$, $Y_2$, and $Y_3$.
- Within imputation cell, joint hot deck imputation was used.

**Table 1** Point estimation

| Parameter | Method | Mean | Std Var. |
|---|---|---|---|
| $\theta_1$ | Complete Data | 1.00 | 100 |
| | FHDI | 1.00 | 135 |
| $\theta_2$ | Complete Data | 2.00 | 100 |
| | FHDI | 2.00 | 142 |
| $\theta_3$ | Complete Data | 4.00 | 100 |
| | FHDI | 4.00 | 132 |
| $\theta_4$ | Complete Data | 0.34 | 100 |
| Proportion | FHDI | 0.33 | 140 |
| $\theta_5$ | Complete Data | 4.00 | 100 |
| Domain Mean | FHDI | 4.00 | 97 |

**Table 2** Variance estimation

| Parameter | Relative Bias (%) |
|---|---|
| $V(\hat{\theta}_1)$ | 6.69 |
| $V(\hat{\theta}_2)$ | -0.79 |
| $V(\hat{\theta}_3)$ | 4.61 |
| $V(\hat{\theta}_4)$ | 3.82 |
| $V(\hat{\theta}_5)$ | 6.50 |

# Conclusion

- Fractional hot deck imputation is considered.
- Categorical data transformation was used for approximation.
  - Does not rely on parametric model assumptions.
  - Two-stage imputation
    1. Stage 1: Imputation of cells (using conditional cell probability)
    2. Stage 2: Joint hot deck imputation within imputation cell.
- Replication-based approach for imputation variance estimation.
- Useful for general-purpose estimation, including domain estimation.
- To be implemented in SAS: Proc SurveyImpute