Two-phase sampling approach to fractional hot deck imputation

Jongho Im[1], Jae-Kwang Kim[1] and Wayne A. Fuller[1]

Abstract

Hot deck imputation is popular for handling item nonresponse in survey sampling. In hot deck imputation, imputed values are taken from the respondents in the same imputation cell, where imputation cells are used to approximate the imputation model. We extend the fractional hot deck imputation of Kim and Fuller (2004) to the case where the imputation cells are not defined in advance. The proposed method of fractional hot deck imputation is performed in two steps and has a structure similar to that of two-phase systematic sampling. The proposed hot deck imputation method is applicable to multivariate missing data. A replication method is used for variance estimation. Results from two simulation studies are presented.

Key Words: Cell mean model, Item nonresponse, EM algorithm, Multivariate missing, Replication variance estimation.

1. Introduction

Nonresponse is frequently encountered in survey sampling. Unit nonresponse and item nonresponse are two major types of nonresponse (Kalton and Kasprzyk,1986). While weighting adjustment is commonly used to compensate for unit nonresponse, imputation is preferred to handle item nonresponse. Haziza (2009) provides a comprehensive overview of imputation methods.

In hot deck imputation, the imputed values are real observations taken from the respondents in the same sample. Hot deck imputation is popular because it does not create artificial values and does not rely on strong model assumptions. In hot deck imputation, creating imputation cells to achieve homogeneity within imputation cells is critical. In Brick and Kalton (1996), all auxiliary variables are treated as categorical and imputation cells are formed as a combination of those categorized auxiliary variables. A nearest-neighbor imputation approach that uses a metric distance of auxiliary variables to find the set of donors has been used by Cotton (1991), Rancourt, Särndal and Lee (1994), and Chen and Shao (2000). Haziza and Beaumont (2007) used the score estimated by the regression of response on the auxiliary variables to create imputation cells. Rubin and Schenker (1986) proposed approximate Bayesian bootstrap (ABB) imputation as a hot deck approach to multiple imputation.

Variance estimation after hot deck imputation is a challenging problem because it is well known that naive approach of treating imputed values as if observed underestimates the true variance. Rubin (1987) proposed multiple imputation as a general tool for inference with imputed data. In multiple imputation, $M(>1)$, imputed estimates are created for each missing item and then the imputation values are used for variance estimation.

Fractional imputation proposed by Kalton and Kish (1984), and investigated by Kim and Fuller (2004), is a way of achieving efficient hot deck imputation. As in multiple imputation, M imputed values are generated for each missing value, but a single data set is created after fractional imputation. Fractional weights are assigned to the imputed values

[1]Department of Statistics, Iowa State University, Ames, U.S.A.

and replication methods are used for variance estimation. Kim and Fuller (2004) and Fuller and Kim (2005) describe some properties of fractional hot deck imputation and discuss variance estimation. Imputation cells are pre-determined and the determination of imputation cells is not discussed in the fractional hot deck imputation of Kim and Fuller (2004).

In this paper, we extend fractional hot deck imputation in two ways. First, instead of assuming the imputation cells to be given, we allow multiple cells for each missing item. The multiple cells can be understood to be a nonparametric approximation of the true model by a finite mixture model. The implementation of fractional hot deck imputation under the finite mixture model is made through a two-phase systematic sampling mechanism. Second, the proposed method is applied to multivariate missing data with arbitrary missing patterns.

In Section 2, the basic setup is introduced. The proposed two-phase fractional imputation and variance estimator are discussed for the univariate case in Section 3. In Section 4, the proposed method is extended to the case of multivariate missing data. Results from two limited simulation studies are presented in Section 5, with concluding remarks in Section 6.

2. Basic setup: univariate missing case

Suppose that we have a finite population of size N, indexed by $U = \{1, 2, \cdots, N\}$, and let A be the index set for the units in the sample selected by a probability sampling mechanism. Let A be partitioned into G groups based on the auxiliary information x, where x takes values on $\{1, \cdots, G\}$. Thus, we can write $A = A_1 \cup \cdots \cup A_G$. In addition to x, we collect y and z where y is the study variable and z is another categorical variable that takes values on $\{1, \cdots, H\}$. The cross classification of x and z forms imputation cells and we assume that

$$y_i \mid (x_i = g, z_i = h) \sim ii(\mu_{gh}, \sigma^2_{gh}), \quad i \in U, \tag{1}$$

for some μ_{gh} and $\sigma^2_{gh} > 0$, where $\sim ii$ denotes independently and identically distributed. Here, x_i is always observed but (y_i, z_i) is subject to missingness. Define $\delta_i = 1$ if (y_i, z_i) is observed and $\delta_i = 0$ otherwise. From unit responses, A can be re-partitioned into $A_R = \{j \in A; \delta_j = 1\}$, $A_M = \{j \in A; \delta_j = 0\}$ with $A = A_M \cup A_R$. Also, A_g can be sub-partitioned into $A_{Rg} = \{j \in A_g; \delta_j = 1\}$ and $A_{Mg} = \{j \in A_g; \delta_j = 0\}$. Let n_{Rg} and n_{Mg} be respectively the size of A_{Rg} and A_{Mg}.

We assume that the response mechanism is missing at random (MAR) in the sense that δ is conditionally independent of (y, z) given x. That is,

$$f(y, z \mid x, \delta) = f(y, z \mid x). \tag{2}$$

The MAR condition (2) implies that model (1) also holds for the responding units. That is,

$$y_i \mid (x_i = g, z_i = h, \delta_i = 1) \sim ii(\mu_{gh}, \sigma^2_{gh}). \tag{3}$$

We now consider a hot deck imputation estimator of $Y_N = \sum_{i=1}^N y_i$ under nonresponse. By the condition (2),

$$f(y \mid x, \delta = 1) \;\; = \;\; \sum_{h=1}^{H} P(z = h \mid x, \delta = 1) f(y \mid x, z = h, \delta = 1). \tag{4}$$

Expression (4) takes the form of a finite mixture model. Let $\pi_{h|g} = P(z = h \mid x = g, \delta = 1)$ be the conditional probability of $z = h$ given $x = g$. The vector (x, z) defines the

imputation cell for hot deck imputation. Note that, from (4),

$$E(y_i \mid x_i = g, \delta_i = 1) = \sum_{h=1}^{H} \pi_{h|g} E(y_i \mid x_i = g, z_i = h, \delta_i = 1).$$

Thus, if $\pi_{h|g}$ is known, we use all respondents in the cell to estimate $E(y_i \mid x_i = g, z_i = h)$ to get

$$\hat{Y}_{FEFI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1 - \delta_i) \sum_{h=1}^{H} \pi_{h|g} \hat{\mu}_{gh} \right\}, \tag{5}$$

where

$$\hat{\mu}_{gh} = \frac{\sum_{j \in A} w_j \delta_j a_{jgh} y_j}{\sum_{j \in A} w_j \delta_j a_{jgh}},$$

with $a_{jgh} = 1$ if $x_j = g$ and $z_j = h$, and $a_{jgh} = 0$ otherwise. The estimator of (5) uses all observed values as donors in the imputation cell and is called the fully efficient fractional efficient (FEFI) estimator (Kim and Fuller, 2004).

3. Fractional hot deck imputation: univariate missing case

We now propose a new fractional hot deck imputation (FHDI) procedure that does not require imputation cell information be given in advance. Given the finite mixture model in (4), the imputed values are taken from the imputation cells with propbability proportional to the conditional cell probabilities, $\pi_{h|g}$. In practice, the cell probabilities $\pi_{h|g}$ are unknown and need to be estimated.

The proposed fractional hot deck imputation is similar in spirit to two-phase sampling for stratification (Rao, 1973; Kim, Navarro, and Fuller, 2006). In phase one, the cells are determined and the cell probabilities $\pi_{h|g}$ are estimated. In phase two, M donors are selected in each imputation cell.

The $\pi_{h|g}$ are estimated so that $\sum_{h=1}^{H} \hat{\pi}_{h|g} = 1$ for each group g. Using the definition $\pi_{h|g} = Pr(z_i = h \mid x_i = g, \delta_i = 1)$, an estimator of $\pi_{h|g}$ is

$$\hat{\pi}_{h|g} = \frac{\sum_{j \in A} w_j \delta_j a_{jgh}}{\sum_{j \in A} w_j \delta_j a_{jg}}, \tag{6}$$

where $a_{jg} = \sum_{h=1}^{H} a_{jgh}$. Thus, the FEFI estimator of (5) can be rewritten as

$$\begin{aligned} \hat{Y}_{FEFI} &= \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1 - \delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g} \hat{\mu}_{gh} \right\}, \\ &= \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1 - \delta_i) \sum_{j \in A} w^*_{ij,FEFI} y_j \right\}, \end{aligned} \tag{7}$$

where $w^*_{ij,FEFI} = \sum_{h=1}^{H} \hat{\pi}_{h|g} \{ w_j \delta_j a_{jgh} / \sum_{l \in A} \delta_l w_l a_{lgh} \}$ is the fractional weights of the j-th donor for the i-th recipient.

Given M imputed values selected for each recipient, the two-phase fractional imputation (FI) estimator of Y_N is defined as

$$\begin{aligned}\hat{Y}_{FI} &= \sum_{g=1}^{G}\sum_{i\in A_g} w_i \left\{ \delta_i y_i + (1-\delta_i)\sum_{h=1}^{H} \hat{\pi}_{h|g}\bar{y}_i^* \right\} \\ &= \sum_{g=1}^{G}\sum_{i\in A_g} w_i \left\{ \delta_i y_i + (1-\delta_i)\sum_{j=1}^{M} w_{ij}^* y_i^{*(j)} \right\}, \end{aligned} \tag{8}$$

where $y_i^{*(j)}$ is the j-th imputed value of y_i, $\bar{y}_i^* = M^{-1}\sum_{j=1}^{M} y_i^{*(j)}$ is a mean of imputed values, and w_{ij}^* are the fractional weights for the FI estimator.

Note that the FI estimator can be also expressed in terms of the FEFI estimator,

$$\begin{aligned}\hat{Y}_{FI} &= \hat{Y}_{FEFI} + (\hat{Y}_{FI} - \hat{Y}_{FEFI}) \\ &= \hat{Y}_{FEFI} + \sum_{g=1}^{G}\sum_{i\in A_{Mg}} w_i(\bar{y}_i^* - \hat{\mu}_g), \end{aligned} \tag{9}$$

where $\hat{\mu}_g = \sum_{j\in A_{Rg}} w_j y_j / \sum_{j\in A_{Rg}} w_j$. While the fully efficient fractional imputation estimator $\hat{Y}_{FEFI}$ has no variance due to the selection of donors, the fractional estimator $\hat{Y}_{FI}$ has additional variance due to the donor selection procedure. Theorem 1 presents some asymptotic properties of the FI estimator.

Theorem 1 *Let the fractional hot deck imputation estimator $\hat{Y}_{FI}$ in (8) be constructed using the two-phase systematic pps sampling.*

(A1) A sequence of probability samples is drawn from a sequence of finite populations (Fuller, 2009) and $\hat{Y}_n = \sum_{i\in A} w_i y_i$ is design-unbiased for Y_N, where w_i is the inverse of the selection probability.

(A2) The cell mean model (1) and the MAR condition (2) hold for the sequence of populations and samples.

(A3) Let U_g, $g = 1, \ldots, G$, be subsets of the finite population with size of N_g, where N_g is fixed, and

$$\left(\hat{N}_g - N_g, \hat{N}_{Rg} - N_{Rg}, \hat{Y}_{Rg} - Y_{Rg}\right) = O_p(n^{-1/2}N),$$

where $(\hat{N}_g, \hat{N}_{Rg}, \hat{Y}_{Rg}) = \sum_{i\in A_g} w_i(1, \delta_i, \delta_i y_i)$ and $(N_g, N_{Rg}, Y_{Rg}) = \sum_{i\in U_g}(1, \delta_i, \delta_i y_i)$.

(A5) The cell mean estimator, $\hat{\mu}_g$, satisfies $\hat{\mu}_g - \mu_g = O_p(n^{-1/2})$, where $\hat{\mu}_g = Y_{Rg}/N_{Rg}$ and $\mu_g = \sum_{h=1}^{H} \pi_{h|g}\mu_{gh}$.

Then,

$$\hat{Y}_{FI} = \tilde{Y}_{FI} + o_p(n^{-1/2}N), \tag{10}$$

and

$$E(\tilde{Y}_{FI} - Y_N) = 0, \tag{11}$$

where

$$\tilde{Y}_{FI} = \tilde{Y}_{FEFI} + \sum_{g=1}^{G}\sum_{i\in A_{Mg}} w_i(\bar{y}_i^* - \hat{\mu}_g),$$

and

$$\tilde{Y}_{FEFI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \{\mu_g + R_g^{-1} \delta_i (y_i - \mu_g)\},$$

with $R_g = N_{Rg}/N_g$. *Also, we have*

$$V(\tilde{Y}_{FI}) = V\left(\tilde{Y}_{FEFI}\right) + E\left\{\sum_{g=1}^{G} \sum_{i \in A_{Mg}} w_i^2 V(\bar{y}_i^* - \hat{\mu}_g \mid A_g)\right\}, \tag{12}$$

and

$$V\left(\tilde{Y}_{FEFI}\right) = V\left(\sum_{g=1}^{G} \sum_{i \in A_g} w_i \mu_g\right) + E\left\{\sum_{g=1}^{G} R_g^{-2} \sum_{i \in A_g} w_i^2 \delta_i (y_i - \mu_g)^2\right\}. \tag{13}$$

See Appendix A for the proof. □

By equation (12), the variance of $\hat{Y}_{FI}$ is the variance of $\hat{Y}_{FEFI}$ plus the variance of the mean of sample donors as an estimator of the cell mean.

Our strategy is to obtain an approximation of FEFI using systematic PPS sampling. Note that, for recipient $i \in A_{Mg}$, we have n_{Rg} FEFI donors with the fractional weights $w_{ij,FEFI}^*$. Thus, M imputed values for y_i can be systematically selected from the donors in A_{Rg} with probability proportional to $w_{ij,FEFI}^*$. The detailed procedure is given in Appendix A. The efficiency of the procedures will depend on the efficiency of the sampling scheme used to select donors.

We propose to estimate the variance by a variance estimator that approximate the variance estimator for the FEFI estimator. Recall that the donors are ordered on the y-variable. A method to construct jackknife replicates is:

(V1) Delete unit k. If $k \in A_M$, then the w_{ij}^* are not changed for $i \in A_{Mg}$.

(V2) For $i \in A_{Mg}$, $k \in A_{Rg}$, and r the closest integer to k, then the w_{ij}^* for replicate k are,

$$w_{ij}^{*(k)} = \begin{cases} w_{ij}^* - w_{ij,FEFI}^* & \text{if } j = r \\ w_{ij}^* + (w_{rj,FEFI}^*) \frac{w_{ij,FEFI}^*}{\sum_{j \neq s} w_{ij,FEFI}^*} & \text{if } j \neq r \\ w_{ij}^* & \text{otherwise.} \end{cases}$$

For each $k \in A_{Rg}$, we identify the nearest FI donor to the deleted element among $\{1, \cdots, M\}$. For example, if the FEFI donor set has 20 donors $\{1, 2, \cdots, 20\}$ ordered on $[j]$ and we have FI with $M = 5$ donors, $\{2, 7, 12, 18, 20\}$, then the nearest FI donor for $k = 3$ is $r = 2$.

Once $w_{ij}^{*(k)}$ are obtained for each $k = 1, \ldots, n$, then a jackknife variance estimator is

$$\hat{V}(\hat{Y}_{FI}) = \sum_{k=1}^{n} c_k (\hat{Y}_{FI}^{(k)} - \hat{Y}_{FI})^2, \tag{14}$$

where

$$\hat{Y}_{FI}^{(k)} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i^{(k)} \left\{\delta_i y_i + (1 - \delta_i) \sum_{j=1}^{M} w_{ij}^{*(k)} y_i^{*(j)}\right\}.$$

Note that the variance estimator (14) is biased because the changes in the fractional weights for the deleted unit are only partially reflected by deleting the closest unit. As demonstrated in the simulation study in Section 5, the relative bias in variance estimator decrease as imputation size M increases.

4. Extension to Multivariate missing data

We now extend the proposed method in Section 3 to multivariate missing case, $\mathbf{y} = (y_1, \cdots, y_p)$. For simplicity in description, we present the proposed method for two variables. For each item, assume that we have discretized values of y_p, denoted by z_p, that can be predetermined or approximated using sample quantiles. Assume that z_1 takes values of $\{1, \ldots, Q\}$ and z_2 takes values of $\{1, \ldots, S\}$. Let δ_{pi} be the response indicator function for y_{pi}. If y_{pi} is missing, then z_{pi} is also missing.

Note that $\mathbf{z} = (z_1, z_2)$ can be viewed as (x, z) of univariate missing case, but z_1 and z_2 are now both subject missingness. That is, A cannot be partitioned into subgroups based on z_1 or z_2. However, we can decompose A such that $A = A_R \cup A_M$, where $A_R = \{j \in A; \delta_j = 1\}$ and $A_M = \{j \in A; \delta_j = 0\}$ with $\delta_j = \prod_{l=1}^{p} \delta_{lj}$. A proposed strategy is that missing items for a unit in A_M are imputed using values of a donor in A_R. For example, if y_{1i} and y_{2i} are both missing, then imputed values of $y_{1i}^{*(j)}$ and $y_{2i}^{*(j)}$ should be selected from the same donor j. Thus, A_R is assumed to be non-empty.

Let $(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis})$ be respectively the observation parts and the missing parts of $\mathbf{y}_i$. Similarly, we have $(\mathbf{z}_{i,obs}, \mathbf{z}_{i,mis})$ for $\mathbf{z}$. There are four patterns of δ_i as $(\delta_{1i} = 1, \delta_{2i} = 1)$, $(\delta_{1i} = 1, \delta_{2i} = 0)$, $(\delta_{1i} = 0, \delta_{2i} = 1)$, and $(\delta_{1i} = 0, \delta_{2i} = 0)$. Thus, we have $(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}) = (y_{2i}, y_{1i} =?)$ and $(\mathbf{z}_{i,obs}, \mathbf{z}_{i,mis}) = (z_{2i}, z_{1i} =?)$ for $(\delta_{1i} = 0, \delta_{2i} = 1)$ pattern, where ? denotes missing value. Similarly, we can identify $\mathbf{y} = (\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis})$ and $\mathbf{z} = (\mathbf{z}_{i,obs}, \mathbf{z}_{i,mis})$ for other patterns.

We assume that the cell mean model (3) holds for cells determined by $\mathbf{z}$,

$$y_{pi} \mid (z_1 = q, z_2 = s) \sim ii(\mu_{pqs}, \sigma^2_{pqs}). \tag{15}$$

Once $\mathbf{z}_{i,mis}$ are imputed, then the conditional distribution of $f(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs})$ can be approximated by

$$f(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs}) \cong \sum_{\mathbf{z}^*_{i,mis}} p(\mathbf{z}^*_{i,mis} \mid \mathbf{z}_{i,obs}) f(\mathbf{y}^*_{i,mis} \mid \mathbf{z}_{i,obs}, \mathbf{z}^*_{i,mis}), \tag{16}$$

where $p(\mathbf{z}^*_{i,mis} \mid \mathbf{z}_{i,obs})$ is the conditional cell probability of $\mathbf{z}^*_{i,mis}$ given $\mathbf{z}_{i,obs}$.

The selection of donors is similar to the univariate missing case in the sense that the approximation (16) has the same mixture model structure of (4) under the MAR assumption. The multivariate version of two-phase systematic pps sampling is as follows:

[Phase 1]: Estimation of cell probabilities

First step for the multivariate hot deck imputation is to estimate the joint cell probabilities $p(\mathbf{z})$, where $p(\mathbf{z})$ is a cell probability for a particular value of $\mathbf{z}$. Since we have missing items on $\mathbf{z}_{i,mis}$, we cannot directly estimate cell probabilities as in the univariate missing case. We use a modified EM algorithm. The procedure avoids producing positive probabilities for structural zeros. See Appendix C for a description of the modified EM algorithm.

To estimate the conditional cell probabilities, we consider a subdivision of A_M based on observed vectors of $\mathbf{z}$. From the observed $\mathbf{z}$ of recipients, A_M can be sub-partitioned into $G = Q + S + 1$ groups, denoted by $\mathbf{z}_1, \ldots, \mathbf{z}_G$, corresponding $\{(1, \ldots, Q), ?)\}$, $\{?, (1, \ldots, S)\}$, and $(?, ?)$. Thus, any nonresponding unit $i \in A_M$ belongs to one subgroup, $A_{Mg} = \{j \in A_M; \mathbf{z}_{j,obs} = \mathbf{z}_{g,obs}\}$, $g = 1, \ldots, G$. We also define $A_{Rg} = \{j \in A_R; \mathbf{z}_{j,obs} = \mathbf{z}_{g,obs}\}$, $g = 1, \ldots, G$. A responding unit $i \in A_R$ can belong to multiple subgroups. For example, $\mathbf{z}_i = (1, 1)$ can be an element of subgroups for both $\mathbf{z}_g = (1, ?)$ and $\mathbf{z}_g = (?, 1)$.

Now, let H_g be the size of all possible vectors for $\mathbf{z}_{g,mis}$. For $i \in A_{Mg}$, if $z^*_{g,mis}$ are imputed for $z_{g,mis}$, then an estimated conditional cell probability $\hat{p}(\mathbf{z}^*_{g,mis} \mid \mathbf{z}_{g,obs})$ is

$$\hat{\pi}_{h|g} = \hat{p}(\mathbf{z}^{*(h)}_g) / \sum_{h=1}^{H_g} \hat{p}(\mathbf{z}^{*(h)}_g), \tag{17}$$

where $\mathbf{z}^{*(h)}_g = (\mathbf{z}_{g,obs}, \mathbf{z}^{*(h)}_{g,mis})$. Then, the FEFI estimator of $Y_p = \sum_{i=1}^N y_{pi}$ is

$$\begin{aligned} \hat{Y}_{p,FEFI} &= \sum_{i\in A} w_i \left\{ \delta_{pi} y_{pi} + \sum_{g=1}^{G} (1-\delta_{pi}) a_{ig} \sum_{h=1}^{H_g} \hat{\pi}_{h|g} \hat{\mu}_{gh} \right\} \\ &= \sum_{i\in A} w_i \left\{ \delta_{pi} y_{pi} + (1-\delta_{pi}) \sum_{j\in A} w^*_{ij,FEFI} y_{pj} \right\} \end{aligned}$$

where $a_{jgh} = 1$ if $(\mathbf{z}_{j,obs}, \mathbf{z}_{j,mis}) = (\mathbf{z}_{g,obs}, \mathbf{z}^{*(h)}_{g,mis})$ and 0 otherwise, $a_{jg} = \sum_{h=1}^{Hg} a_{jgh}$, $w^*_{ij,FEFI} = \sum_{g=1}^{G} a_{ig} \sum_{h=1}^{H} \hat{\pi}_{h|g} \{ w_j \delta_j a_{jgh} / \sum_{l\in A} w_l \delta_l a_{lgh} \}$, and

$$\hat{\mu}_{pgh} = \frac{\sum_{j\in A} w_j \delta_j a_{jgh} y_{pj}}{\sum_{j\in A} w_j \delta_j a_{jgh}}.$$

[Phase 2]: Systematic PPS sampling for missing $\mathbf{y}_{mis}$

The fractional hot deck imputation for multivariate case can be implemented in a manner similar to the univariate case. The (S1) procedure in Appendix A needs to be replaced with the following (M1):

(M1) Let n_{Rg} be the number of FEFI donors in A_{Rg}. Then, n_{Rg} FEFI donors are ordered with $\mathbf{y}$ values by the half-ascending and half-descending order as in the univariate case. The sorting method depends on the number of missing items in $\mathbf{z}_{g,mis}$,

(a) Single missing item: the FEFI donors are sorted based on y_p values corresponding to the missing item.

(b) Multiple missing items: the FEFI donors are first sorted by values of z that has the highest response rate among missing items. After then, the FEFI donors are sequentially sorted by values of z in order of item response rates. Note that, if we use y instead of z in sorting of the FEFI donors, the final order only depends on the missing item that has the lowest response rate.

Once M donors are selected from the FEFI donors with probability proportional to $w^*_{ij,FEFI}$ for each recipient, the FI estimator of Y_p is

$$\begin{aligned} \hat{Y}_{p,FI} &= \sum_{i\in A} w_i \left\{ \delta_{pi} y_{pi} + \sum_{g=1}^{G} (1-\delta_{pi}) a_{ig} \sum_{h=1}^{H_g} \hat{\pi}_{h|g} \bar{y}^*_{pi} \right\} \\ &= \sum_{i\in A} w_i \left\{ \delta_{pi} y_{pi} + (1-\delta_{pi}) \sum_{j=1}^{M} w^*_{ij} y^{*(j)}_{pi} \right\} \end{aligned}$$

where $y^{*(j)}_{pi}$ is the y_{pj} of j-th donor for y_{pi}, $\bar{y}^*_{pi} = M^{-1} \sum_{j=1}^M y^{*(j)}_{pi}$ is a mean of imputed values, and $w^*_{ij} = M^{-1}$.

For variance estimation, we first calculate $w^{*(k)}_{ij}$ based on $(V1)$ and $(V2)$ in Section 3 and then compute the jackknife variance estimate from the formula in (14).

5. Simulation Study

5.1 Univariate missing case

To check the performance of the proposed method in the univariate case, $Y_i = (Y_{1i}, Y_{2i})$, $i = 1, \cdots, n$ are randomly generated from

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

where e_2 is independent of Y_1 and is generated from a standard normal distribution. Here, Y_1 is fully observed observed but Y_2 is subject missingness with $\delta \sim Bernoulli(0.7)$. Thus, Y_1 plays the role of x in Section 3. In the simulation, B=5,000 Monte Carlo samples are generated with size of $n = 300$.

To implement the fractional hot deck imputation, Y_1 and Y_2 are categorized into $\tilde{Y}_1$ and $\tilde{Y}_2$ that respectively play roles of x and z of Section 2. The auxiliary variable ,Y_1, is categorized into five groups and the study variable, Y_2, is partitioned into two groups based on the sample quantiles of the respondents. For example, observations with y_2 values less than the median belong to group 1 (i.e. $\tilde{y}_2 = 1$).

For each recipient, $M = 10$ and $M = 20$ donors are respectively selected using the systematic sampling with probability proportional to the fractional weights of the FEFI donors. If M is no greater than n_{Rg}, then we select all FEFI donors and use FEFI donors' fractional weights as the fractional weights for the FI estimator.

We consider five parameters: $\theta_1 = E(Y_2)$, $\theta_2 = P(Y_2 < 2)$, $\theta_3 = E(Y_2 \mid D = 1)$ with $D \sim Bernoulli(0.3)$, θ_4 is the slope of regression of Y_2 on Y_1 and θ_5 is the correlation between Y_1 and Y_2. The five parameters are estimated using the FEFI estimator and the FI estimator. For θ_4 and θ_5, the FEFI or the FI estimator is

$$\hat{\theta}_4 = \frac{\sum_{i\in A}\{\delta_i(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}^*_{2I}) + (1-\delta_i)\sum_{j\in A}\delta_j w^*_{ij}(y_{1i} - \bar{y}_1)(y_{2j} - \bar{y}^*_{2I})\}}{\sum_{i\in A}(y_{1i} - \bar{y}_1)^2},$$

and

$$\hat{\theta}_5 = \frac{\sum_{i\in A}\{\delta_i(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}^*_{2I}) + (1-\delta_i)\sum_{j\in A}\delta_j w^*_{ij}(y_{1i} - \bar{y}_1)(y_{2j} - \bar{y}^*_{2I})\}}{\{\sum_{i\in A}(y_{1i} - \bar{y}_1)^2\}^{1/2}[\sum_{i\in A}\{\delta_i(y_{2i} - \bar{y}^*_{2I})^2 + (1-\delta_i)\sum_{j\in A}\delta_j w^*_{ij}(y_{2j} - \bar{y}^*_{2I})^2\}]^{1/2}},$$

where $\bar{y}^*_{2I}$ is a mean of imputed samples in y_2. In addition to point estimators, we also computed variance estimators using the replication method in Section 3. The jackknife variance estimator in (14) is used to compute variance estimates of the FI estimator.

Table 1 presents the Monte Carlo means, standardized variances of the point estimators and relative bias of variance estimates. All point estimators are nearly unbiased. Slight biases in estimation of regression slope and correlation are due to the discrete approximation. The variances are reported with respect to variances of the FEFI estimators. The FI esimators are as efficient as the FEFI estimators. For larger imputation size, we have smaller relative bias in variance estimates.

5.2 Multivariate case

Now we extend the proposed method to a multivariate missing case. We generated $Y_i = (Y_{1i}, Y_{2i})$, $i = 1, \cdots, n$, from

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

Table 1: Monte Carlo results of mean, standardized variance and relative bias of variance estimators for the univariate missing case

Parameter	Estimator	Bias of $\hat{\theta}$	Std. Var. of $\hat{\theta}$	Rel.Bias (%) of $\hat{V}\{\hat{\theta}\}$
θ_1	FEFI	-0.00019	100.0	
$E(Y_2)$	FI(M=10)	-0.00016	100.0	-8.0
	FI(M=20)	-0.00019	100.0	-6.1
θ_2	FEFI	-0.00009	100.0	
$P(Y_2 < 2)$	FI(M=10)	-0.00008	100.0	-2.7
	FI(M=20)	-0.00009	100.0	-0.8
θ_3	FEFI	0.00125	100.0	
$E(Y_2 \mid D = 1)$	FI(M=10)	0.00113	100.0	-2.7
	FI(M=20)	0.00124	100.0	-1.5
θ_4	FEFI	-0.01201	100.0	
Slope	FI(M=10)	-0.01202	100.0	-6.5
	FI(M=20)	-0.01198	100.0	-3.1
θ_5	FEFI	-0.00599	100.0	
Corr(Y_1, Y_2)	FI(M=10)	-0.00599	100.0	-7.9
	FI(M=20)	-0.00598	100.0	-3.9

where e_2 and e_3 are independently generated from a standard distribution. We generated $\delta_{li} \sim Bernoulli(p_l)$ independently for each Y_l, $l = 1, 2, 3$, with $p_1 = 0.5$, $p_2 = 0.7$ and $p_3 = 0.9$ so that all variables are subject to missingness.

Each variable is firstly categorized into three groups and then collapsed so that A_{Rg} has at least two elements. We select $M = 10$ and $M = 20$ donors for each recipient using systematic sampling with probability proportional to the fractional weights of the FEFI donors. If $M \geq n_{Rg}$, then we select all possible donors and assign the fractional weights of the FEFI estimator as the fractional weights of the FI estimator. We generate $B = 5,000$ Monte Carlo samples with size of $n = 500$.

We computed estimators of $\theta_1 = E(Y_1)$, $\theta_2 = E(Y_2)$, $\theta_3 = E(Y_3)$, $\theta_4 = P(Y_1 < 1, Y_2 < 2)$ and $\theta_5 = E(Y_2 \mid D = 1)$ with $D \sim Bernoulli(0.3)$. For variance estimation of the FI estimator, we used the jackknife estimator with formula (14).

Table 2 presents the Monte Carlo means, standardized variances of the point estimators and relative bias of variance estimates. All estimators are nearly unbiased and the proposed FEFI and FI estimator perform well in this simulation. The variances are reported based on the FEFI estimates. As univariate missing case, the FI estimators are also as efficient as the FEFI estimators. The relative biases of variance estimators are smaller with $M = 20$ than the biases with $M = 10$.

6. Concluding remarks

A fractional hot deck imputation in this paper mimics two-phase systematic sampling in the sense that imputation cells are created and missing items are imputed using the cells. The variance estimator has a potential for bias because the component due to donor selection is ignored.

For multivariate imputation, joint cell probabilities are used to define conditional cell probabilities. The joint distribution of the study vector is approximated by a discrete approximation. The choice for the optimal level of discrete approximation can be viewed

Table 2: Monte Carlo results of mean, standardized variance and relative bias of variance estimators for the multivariate missing case

Parameter	Estimator	Bias of $\hat{\theta}$	Std. Var. of $\hat{\theta}$	Rel.Bias (%) of $\hat{V}\{\hat{\theta}\}$
θ_1	FEFI	-0.00053	100.0	
$E(Y_1)$	FI(M=10)	-0.00052	100.0	-6.4
	FI(M=20)	-0.00054	100.0	-3.5
θ_2	FEFI	-0.00017	100.0	
$E(Y_2)$	FI(M=10)	-0.00015	100.0	-10.2
	FI(M=20)	-0.00016	100.0	-5.7
θ_3	FEFI	-0.00156	100.0	
$E(Y_3)$	FI(M=10)	-0.00156	100.0	-8.2
	FI(M=20)	-0.00156	100.0	-6.9
θ_4	FEFI	-0.00375	100.0	
$E(Y_1 < 1, Y_2 < 2)$	FI(M=10)	-0.00376	100.0	-5.3
	FI(M=20)	-0.00375	100.0	0.3
θ_5	FEFI	0.00071	100.0	
$E(Y_2 \mid D = 1)$	FI(M=10)	0.00065	100.0	-3.6
	FI(M=20)	0.00077	100.0	-2.0

as bandwidth selection for a nonparametric procedure. A modified EM algorithm is introduced for computation of joint cell probabilities.

One desirable feature of the proposed method is that the covariance structure of multivariate variables is retained after imputation because imputed values are jointly generated and are selected to mimic distribution of variables as closely possible. An efficient sampling algorithm such as systematic PPS sampling is required. While the proposed FI estimator is nearly as efficient as the FEFI estimator, the size of the finally imputed data set will be relatively small compared to the use of FEFI. An **R** software package of the proposed method is under development.

Appendix

A. Systematic PPS sampling procedure

(S1) Sort n_{Rg} FEFI donors in terms of y values by the half-ascending and half-descending order. For example, $\{1, 2, \ldots, 10, 11\}$ is sorted as follows: $1, 3, 5, 7, 9, 11, 10, 8, 6, 4, 2$. Let $[j]$,$j = 1, \ldots, n_{Rg}$, be the j-th sorted unit in A_{Rg}.

(S2) Construct the interval of $(L_{[j]}, U_{[j]})$ for the systematic pps sampling.

  (a) Set $j = 1$ and $L_{[1]} = 0$.

  (b) For current j, $U_{[j]} = L_{[j]} + M \times w^*_{ij,FEFI}$

  (c) Set $j = j + 1$ and $L_{[j]} = U_{[j-1]}$ and go to step (b) until $j = n_{Rg}$.

(S3) Let $(RN)_g$ be a random number generated from $U(0, 1)$. For each $i \in A_{Mg}$, we select M donors as follows: For $l = 1, \ldots, M$, if

$$L_{[j]} \leq \frac{(RN)_g + (i-1)}{n_{Mg}} + (l-1) \leq U_{[j]}$$

for some j, then $y_{[j]}$ be the l-th imputed value for unit i.

B. Proof of Theorem 1

Before we prove Theorem 1, assume that δ_i $(i = 1, \ldots, N)$ is extended to the entire population and assumed to be independent random variable. This extension has been discussed by Fay (1991) and used by Rao and Shao (1992).

First, we rewrite the FEFI estimator in (7),

$$\begin{aligned}
\hat{Y}_{FEFI} &= \sum_{g=1}^{G}\sum_{i\in A_g} w_i\delta_i y_i + \sum_{g=1}^{G}\sum_{i\in A_g} w_i(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}\hat{\mu}_{gh} \\
&= \sum_{g=1}^{G}\sum_{i\in A_g} w_i\delta_i \sum_{h=1}^{H}\hat{\pi}_{h|g}\hat{\mu}_{gh} + \sum_{g=1}^{G}\sum_{i\in A_g} w_i(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}\hat{\mu}_{gh} \\
&= \sum_{g=1}^{G}\sum_{i\in A_g} w_i \sum_{h=1}^{H}\hat{\pi}_{h|g}\hat{\mu}_{gh} \\
&= \sum_{g=1}^{G}\hat{N}_g(\hat{Y}_{Rg}/\hat{N}_{Rg}) \qquad \text{(B.1)}
\end{aligned}$$

Now, applying Taylor expansion on the $\hat{Y}_{FEFI}$ defined in (B.1), we have

$$\begin{aligned}
\hat{Y}_{FEFI} &= \sum_{g=1}^{G}\frac{N_g}{N_{Rg}}Y_{Rg} + \sum_{g=1}^{G}\frac{N_g}{N_{Rg}}(\hat{Y}_{Rg} - Y_{Rg}) \\
&\quad + \sum_{g=1}^{G}\frac{Y_{Rg}}{N_{Rg}}(\hat{N}_g - N_g) - \sum_{g=1}^{G}\frac{Y_{Rg}N_g}{N_{Rg}^2}(\hat{N}_{Rg} - N_{Rg}) + S_n + G_n, \qquad \text{(B.2)}
\end{aligned}$$

where

$$\begin{aligned}
S_n &= \frac{1}{N_{Rg}}(\hat{Y}_{Rg} - Y_{Rg})(\hat{N}_g - N_g) - \frac{N_g}{N_{Rg}^2}(\hat{Y}_{Rg} - Y_{Rg})(\hat{N}_{Rg} - N_{Rg}) \\
&- \frac{Y_{Rg}}{N_{Rg}^2}(\hat{N}_{Rg} - N_{Rg})(\hat{N}_g - N_g) + \frac{Y_{Rg}N_g}{N_{Rg}^3}(\hat{N}_{Rg} - N_{Rg})^2,
\end{aligned}$$

and G_n is a remainder term.

From the assumption $(A4)$, S_n has the order of $O_p(n^{-1}N)$. Thus, by the assumption $(A4)$ and $(A5)$, (B.2) can be expressed with

$$\hat{Y}_{FEFI} = \sum_{g=1}^{G}\sum_{i\in A_g} w_i\{y_i + (R_g^{-1}\delta_i - 1)(y_i - \mu_g)\} + o_p(n^{-1/2}N), \qquad \text{(B.3)}$$

where, $R_g = N_{Rg}/N_g$ and $\mu_g = \sum_{h=1}^{H}\pi_{h|g}\mu_{gh}$ for $i \in U_g$. Henceforth, we define $\tilde{Y}_{FEFI} = \sum_{g=1}^{G}\sum_{i\in A_g} w_i\gamma_{ig}$ with $\gamma_{ig} = y_i + (R_g^{-1}\delta_i - 1)(y_i - \mu_g)$.

Thus, from (9) and $\tilde{Y}_{FEFI}$, we have the result (10),

$$\hat{Y}_{FI} = \tilde{Y}_{FI} + o_p(n^{-1/2}N), \qquad \text{(B.4)}$$

where $\tilde{Y}_{FI} = \tilde{Y}_{FEFI} + \sum_{g=1}^{G}\sum_{i\in A_{Mg}} w_i(\bar{y}_i^* - \hat{\mu}_g)$.

Let $E_I(\cdot)$ be an expectation on imputation mechanism, we have

$$E_I(\tilde{Y}_{FI}) = \tilde{Y}_{FEFI}. \tag{B.5}$$

Thus, to prove (11), it suffices to show that $E(\tilde{Y}_{FEFI} - Y_N) = 0$.

Taking expectation on $\tilde{Y}_{FEFI}$, we have

$$\begin{aligned} E(\tilde{Y}_{FEFI}) &= E\left\{E(\bar{Y}_{FEFI} \mid \mathcal{F}_N)\right\} \\ &= E\left(\sum_{g=1}^{G}\sum_{i\in U_g} y_i\right) + E\left(\sum_{g=1}^{G}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)(y_i - \mu_g)\right) \\ &= E(Y_N) + E\left(\sum_{g=1}^{G}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)(y_i - \mu_g)\right), \end{aligned} \tag{B.6}$$

where $\mathcal{F}_N$ is a set of finite population.

On the other hand,

$$E\left(\sum_{g=1}^{G}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)(y_i - \mu_g)\right) = 0. \tag{B.7}$$

From (B.5), (B.6) and (B.7), we have $E(\tilde{Y}_{FI} - Y_N) = 0$, that is, (11) is established.

We now consider variance of the FI estimator. From expression (B.3), we fist have

$$\begin{aligned} V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\gamma_{ig}\right) &= V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) + V\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i - \mu_g)\right\} \\ &+ \text{Cov}\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g, \sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i - \mu_g)\right\}. \end{aligned} \tag{B.8}$$

For the second term of (B.8), we have

$$V\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i - \mu_g)\right\} = E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2 R_g^{-2}\delta_i(y_i - \mu_g)^2\right\}, \tag{B.9}$$

where the equality comes from $E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i - \mu_g)\right\} = 0$. For the third term of (B.8), we also have

$$\begin{aligned} &\text{Cov}\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g, \sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i - \mu_g)\right\} \\ &= E\left\{\sum_{g=1}^{G}\sum_{i\in U_g} w_i\mu_g R_g^{-1}\delta_i(y_i - \mu_g)\right\} = 0 \end{aligned} \tag{B.10}$$

where the equality comes from $E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i - \mu_g)\right\} = 0$ and the cell mean model in (1).

From (B.8)-(B.10), we write

$$V(\tilde{Y}_{FEFI}) = V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) + E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2 R_g^{-2}\delta_i(y_i - \mu_g)^2\right\} \tag{B.11}$$

Note that the variance of $N^{-2}\tilde{Y}_{FEFI}$ converges to the variance of $N^{-2}\hat{Y}_{FEFI}$ as n goes to infinity such that

$$\begin{aligned} V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\gamma_{ig}\right) &= E\left[\left\{\sum_{g=1}^{G}\hat{R}_g^{-1}\sum_{i\in A_g} w_i\delta_i(y_i-\mu_g)\right\}^2\right] + \text{(Cross product term)} \\ &\quad + E\left[\left\{\sum_{g=1}^{G}(R_g^{-1}-\hat{R}_g^{-1})\sum_{i\in A_g} w_i\delta_i(y_i-\mu_g)\right\}^2\right] \qquad \text{(B.12)} \\ &= V\left(\sum_{g=1}^{G}\hat{R}_g^{-1}\sum_{i\in A_g} w_i\delta_i y_i\right) + O(n^{-3/2}N^2) \\ &= V\left(\hat{Y}_{FEFI}\right) + o(n^{-1}N^2), \qquad \text{(B.13)} \end{aligned}$$

where the second term of (B.12) converges to 0 with order of $O(n^{-2}N^2)$ and the cross product term converges to 0 with order of $O(n^{-3/2}N^2)$ by the condition $(A4)$ and the Schwarz inequality. Thus, from (B.11) and (B.13), we show (13) such that

$$V(\hat{Y}_{FEFI}) = V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) + E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2 R_g^{-2}\delta_i(y_i-\mu_g)^2\right\} + o_p(n^{-1}N^2). \qquad \text{(B.14)}$$

We now write,

$$V\left(\hat{Y}_{FI}-\hat{Y}_{FEFI}\right) = V\left\{E_I\left(\hat{Y}_{FI}-\hat{Y}_{FEFI}\right)\right\} + E\left\{V_I\left(\hat{Y}_{FI}-\hat{Y}_{FEFI}\right)\right\},$$

where $\hat{Y}_{FI}-\hat{Y}_{FEFI} = \sum_{g=1}^{G}\sum_{i\in A_{Mg}} w_i(\bar{y}_i^*-\hat{\mu}_{gh})$ and $V_I(\cdot)$ is a variance on imputation mechanism. On the imputation mechanism,

$$V_I(\hat{Y}_{FI}-\hat{Y}_{FEFI}) = \sum_{g=1}^{G}\sum_{i\in A_{Mg}} w_i^2 V(\bar{y}_i^*-\hat{\mu}_g \mid A_g). \qquad \text{(B.15)}$$

$$E_I(\hat{Y}_{FI}-\hat{Y}_{FEFI}) = 0. \qquad \text{(B.16)}$$

Thus, by the result of (B.15) and (B.16),

$$V\left(\hat{Y}_{FI}-\hat{Y}_{FEFI}\right) = E\left\{\sum_{g=1}^{G}\sum_{i\in A_{Mg}} w_i^2 V(\bar{y}_i^*-\hat{\mu}_g \mid A_g)\right\}. \qquad \text{(B.17)}$$

Also, since $E_I(\hat{Y}_{FI}) = \hat{Y}_{FEFI}$, we have

$$Cov(\hat{Y}_{FI}-\hat{Y}_{FEFI}, \hat{Y}_{FEFI}) = 0 \qquad \text{(B.18)}$$

Therefore, by (B.17) and (B.18), (12) is established

C. Description of the EM algorithm

The EM algorithm is used here in a slightly modified way. For each unit i, the conditional probability of $\mathbf{z}_{i,mis}$ given $\mathbf{z}_{i,obs}$ is computed using the current estimate of the joint probability $\hat{p}(\mathbf{z})$, where $\sum_{\mathbf{z}}\hat{p}(\mathbf{z}) = 1$. This is the E-step of the EM algorithm. The initial conditional probabilities are

$$w_i^{*(h)} = \frac{\hat{p}_0(\mathbf{z}_{i,obs}, \mathbf{z}_{i,mis} = \mathbf{z}_{i,mis}^{*(h)})}{\sum_{h=1}^{H_i}\hat{p}_0(\mathbf{z}_{i,obs}, \mathbf{z}_{i,mis}^{*(h)})}. \qquad \text{(C.1)}$$

where $\hat{p}_0(\mathbf{z})$ is the estimated joint probability computed from the full respondents and $\mathbf{z}_{i,mis}^{*(h)}$ is the h-th imputed vector for the missing missing items of unit $i \in A_M$. Here, H_i denotes the number of imputed vectors in $\mathbf{z}_{i,mis}^{*(h)}$.

The M-step computes the joint probability of particular combination of $\mathbf{z}^* = (\mathbf{z}_{obs}, \mathbf{z}_{mis}^*)$,

$$\hat{p}(\mathbf{z}^*) = (\sum_{i=1}^{n} w_i)^{-1} \sum_{i=1}^{n} \sum_{h=1}^{H_i} w_i w_i^{*(h)} I(\mathbf{z}_{i,obs} = \mathbf{z}_{obs}, \mathbf{z}_{i,mis}^{*(h)} = \mathbf{z}_{mis}^*). \tag{C.2}$$

Equations (C.1) and (C.2) form a set of iterative computations for the EM algorithm. In the iteration, $\hat{p}_0$ is replaced by $\hat{p}$ to compute $w_i^{*(h)}$ and update $\hat{p}(\mathbf{z}^*)$ again until it converges.

REFERENCES

Brick, J.M. and Kalton, G. (1996). "Handling missing data in servey research". *Statistical Methods in Medical Research*, 5, 215–238.

Chen, J. and Shao, J. (2000). "Nearest neighbor imputation for survey data". *Journal of Official Statistics*, 16, 113–132.

Cotton, C. (1991). *Functional description of the Generalized Edit and Imputation System*. Business Survey Methods Division, Statistics Canada.

Fay, R. E. (1991). "A design-based perspective on missing data variance". In *Proceedings of Bureau of the Census Annual Research Conference*, American Statistical Association, 429–440.

Fuller, W.A. (2009). *Sampling Statistics.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Fuller, W.A. and Kim, J.K. (2005). "Hot deck imputation for the response model". *Survey Methodology*, 31, 139–149.

Haziza, D. (2009). "Imputation and inference in the presence of missing data". In *Handbook of Statistics*, volume 29, *Sample Surveys: Theory Methods and Inference*, Edited by C.R. Rao and D. Pfeffermann, 215–246.

Haziza, D. and Beaumont, J.F. (2007). "On the construction of imputation classes in surveys". *International Statistical Review*, 75, 25–43.

Kalton, G. and Kasprzyk, D. (1986). "The treatment of missing survey data". *Survey Methodology*, 12, 1–16.

Kalton, G. and Kish, L. (1984). "Some efficient random imputation methods" *Communications in Statistics*, 13, 1919–1939.

Kim, J.K. and Fuller, W.A. (2004). "Fractional hot deck imputation". *Biometrika*, 91, 559–578.

Kim,J.K., Navarro, A., and Fuller, W.A. (2006). "Replication variance estimation for two-phase stratified sampling". *Journal of American Statistical Association*, 101, 312–320.

Rancourt, E., Särndal, C.E., and Lee, H. (1994). "Estimation of the variance in presence of nearest neighbor imputation". In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 888–893.

Rao, J.N.K. (1973). "On double sampling for stratification and analytical surveys". *Biometrika*, 60, 125–133.

Rao, J.N.K., and Shao, J. (1992). "Jackknife variance estimation with survey data under hot deck imputation". *Biometrika*, 79, 811–822.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*, New York: John Wiley & Sons, Inc.

Rubin, D.B. and Schenker, N. (1986). "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse". *Journal of the American Statistical Association*, 81, 366–374.