Statistical Science
2016, Vol. 31, No. 3, 415–432
DOI: 10.1214/16-STS569

Fractional Imputation in Survey Sampling: A Comparative Review

Shu Yang and Jae Kwang Kim

Abstract. Fractional imputation (FI) is a relatively new method of imputation for handling item nonresponse in survey sampling. In FI, several imputed values with their fractional weights are created for each record with missing items. Each fractional weight represents the conditional probability of the imputed value given the observed data, and the parameters in the conditional probabilities are often computed by an iterative method such as the EM algorithm. The underlying model for FI can be fully parametric, semiparametric or nonparametric, depending on the plausibility of assumptions and the data structure.

In this paper, we give an overview of FI, introduce key ideas and methods to readers who are new to the FI literature, and highlight some new developments. We also provide guidance on practical implementation of FI and valid inferential tools after imputation. We demonstrate the empirical performance of FI with respect to multiple imputation using a pseudo finite population generated from a sample from the Monthly Retail Trade Survey conducted by the US Census Bureau.

Key words and phrases: Item nonresponse, missing at random, Monte Carlo EM, multiple imputation, synthetic imputation.

1. INTRODUCTION

In survey sampling, it is common practice to collect data on a large number of items. Even when a sampled unit responds to the survey, this unit may not respond to some items. In this scenario, imputation can be used to create a complete dataset by filling in missing values with plausible values to facilitate data analyses. Imputation achieves three goals. First, by providing complete data, subsequent analyses are easy to implement and results are consistent among different users. Second, imputation reduces the selection bias associated with only using the respondent set, which may not necessarily represent the original sample. Third, the imputed data can incorporate extra information from outside of the sample so that the resulting analyses are statistically efficient and coherent. Combining information from several surveys or creating synthetic data from planned missingness are cases in point (Schenker and Raghunathan, 2007).

When the imputed dataset is released to the public, it should meet the goal of multiple uses for both planned and unplanned parameters (Haziza, 2009). Consequently, it is a best practice to include some means of estimating the effect of missing data and missing data treatment along with the public use datasets, thus allowing data users to assess the data utility for their analyses. Rubin (1976) proposed multiple imputation (MI) for this purpose. MI replaces each missing data item with several plausible values to reflect the full uncertainty in the prediction of missing data, creating M completed datasets. Several authors (Rubin, 1987; Little and Rubin, 2002; Schafer, 1997) have promoted MI as a standard approach for general-purpose estimation under item nonresponse in survey sampling. MI requires special conditions, called congeniality (Meng, 1994) and self-efficient estimation (Meng and Romero, 2003). Otherwise, as discussed by Kott (1995), Fay (1992, 1996), Binder and

Shu Yang is a postdoc fellow, Department of Biostatistics, Harvard University, 655 Huntington Ave, Boston, Massachusetts 02115, USA (e-mail: shuyang@hsph.harvard.com). Jae Kwang Kim is a professor, Department of Statistics, Iowa State University, Ames, Iowa 50011, USA (e-mail: jkim@iastate.edu).

Sun (1996), Wang and Robins (1998), Nielsen (2003) and Kim et al. (2006), the MI variance estimator is not always consistent. The inconsistency occurs even when the imputation model is correctly specified. This phenomenon occurs when the complete-sample analyses are not self-efficient (Meng and Romero, 2003). Beaumont, Haziza and Bocci (2011) studied the problem of variance estimation for MI and found that the MI variance estimator is considerably biased when the complete-sample estimator is not self-efficient. Yang and Kim (2016b) also found that self-efficiency does not hold for the method of moments estimator.

Fractional imputation (FI) is another effective imputation tool for general-purpose estimation but has the advantage of not requiring the self-efficiency condition. FI was originally proposed by Kalton and Kish (1984) to reduce the variance of single imputation methods by replacing each missing value with several plausible values at differential probabilities reflected through fractional weights. FI can provide a fully efficient estimator of the mean while preserving the distribution of the variable being imputed, therefore allowing for efficient general-purpose estimation. For a univariate y, balanced random imputation proposed by Chauvet, Deville and Haziza (2011) also achieves this goal. Fay (1996), Kim and Fuller (2004), Fuller and Kim (2005), Durrant (2005), Durrant and Skinner (2006) discussed FI as a nonparametric imputation method for descriptive parameters in survey sampling. Kim (2011) and Kim and Yang (2014) presented FI under fully parametric model assumptions.

More generally, FI can serve as a computational tool for implementing the expectation step (E-step) in the EM algorithm (Wei and Tanner, 1990; Kim, 2011). When the conditional expectation in the E-step is not available in a closed form, parametric FI (Kim, 2011) simplifies computation by drawing on importance sampling to obtain the fractional weights and reducing the iterative computation burden over other simulation methods such as Markov Chain Monte Carlo. Kim and Hong (2012) extended parametric FI to a more general class of incomplete data, including measurement error models.

Despite these advantages, FI in applied research has not been widely used perhaps due to the lack of a comprehensive reference and the availability of software. The advantages of FI may come at the cost of an increase in data storage space and computation complexity, compared to MI, since FI may require replication methods for valid variance estimation for general parameters. But if the goal is to achieve statistical efficiency and validity, FI may be preferable. The goal of this paper is to bring more attention to FI by reviewing existing research on FI, introducing key ideas and methods, and highlighting some new developments, mainly in the context of survey sampling. This paper also provides guidance on practical implementation and application of FI.

This paper is organized as follows. Section 2 provides the basic setup, and Section 3 introduces FI under parametric model assumptions. Section 4 discusses a nonparametric approach to FI, specifically in the context of hot deck imputation. Section 5 introduces synthetic data imputation using FI in the context of two-phase sampling. Section 6 deals with practical considerations and variations of FI, including imputation sizes, choices of proposal distributions and doubly robust FI. Section 7 compares FI with MI in terms of statistical efficiency and ability to accommodate informative sampling. Section 8 presents a simulation study based on survey data modeled from the Monthly Retail Trade Survey conducted by the US Census Bureau. Section 9 contains our concluding remarks.

2. BASIC SETUP

Consider a finite population of N units identified by a set of indices $U = \{1, 2, \ldots, N\}$ with N known. The K-dimensional study variable $y_i = (y_{i1}, \ldots, y_{iK})$, associated with each unit i in the population is subject to missingness. We assume that the finite population at hand is a realization from an infinite population, called a *superpopulation*. For the superpopulation, we often postulate a parametric distribution, $f(y; \theta)$, with the parameter $\theta \in \Omega$. We can express the density for the joint distribution of y as

$$f(y;\theta) = f_1(y_1;\theta_1) f_2(y_2 \mid y_1;\theta_2) \cdots \cdot f_K(y_K \mid y_1, \ldots, y_{K-1};\theta_K), \tag{2.1}$$

where θ_k is the parameter in the conditional distribution of y_k given $y_1, \ldots, y_{k-1}$. Let A denote the set of indices for units in the sample selected by a probability sampling mechanism. Each unit is associated with a sampling weight, the inverse of the probability of being selected into the sample, denoted by w_i.

We are interested in estimating η, defined as a (unique) solution to the population estimating equation $\sum_{i=1}^{N} U(\eta; y_i) = 0$. For example, the population mean of y can be obtained by letting $U(\eta; y_i) = \eta - y_i$, the population proportion of y less than a threshold c can be obtained by letting $U(\eta; y_i) = \eta - I_{\{y_i < c\}}$, where I is an indicator function, the population median of y can be obtained by choosing $U(\eta; y_i) = 0.5 - I_{\{y_i < \eta\}}$, and

so on. Under complete response, a design-consistent estimator of η is obtained by solving

$$\sum_{i\in A} w_i U(\eta; y_i) = 0. \tag{2.2}$$

Godambe and Thompson (1986), Binder and Patak (1994) and Rao, Yung and Hidiroglou (2002) have done rigorous investigations of the estimator obtained from (2.2) under complex sampling.

In the presence of missing data, first consider decomposing $y_i = (y_{\text{obs},i}, y_{\text{mis},i})$, where $y_{\text{obs},i}$ and $y_{\text{mis},i}$ are the observed and missing parts of y_i, respectively. We assume that the response mechanism is missing at random (MAR) in the sense of Rubin (1976). That is, the probability of nonresponse only depends on the observed part but not on the missing value itself. Under MAR, a consistent estimator of η can be obtained by solving the conditional estimating equation, given the observed data $y_{\text{obs}} = (y_{\text{obs},1}, \ldots, y_{\text{obs},n})$,

$$\sum_{i\in A} w_i E\{U(\eta; y_i) \mid y_{\text{obs},i}\} = 0, \tag{2.3}$$

where the above conditional expectation is taken with respect to the prediction model (also called the imputation model),

$$f(y_{\text{mis},i} \mid y_{\text{obs},i}; \theta) = \frac{f(y_{\text{obs},i}, y_{\text{mis},i}; \theta)}{\int f(y_{\text{obs},i}, y_{\text{mis},i}; \theta)\, dy_{\text{mis},i}}, \tag{2.4}$$

which depends on the unknown parameter θ. Imputation is therefore a computational tool for computing the conditional expectation in (2.3) for arbitrary choices of the estimating function $U(\eta; y)$. The resulting conditional expectation using imputation can be called the imputed estimating function.

Table 1 presents a summary of Bayesian and frequentist approaches to statistical inference with missing data. In the Bayesian approach, θ is treated as a random variable and the reference distribution is the joint distribution of θ and the latent (missing) data, given the observed data. On the other hand, in the frequentist approach, θ is treated as fixed and the reference distribution is the likelihood distribution of the latent data of a given parameter θ, conditional on the observed data. The learning algorithm for updating parameter estimates from the observed data is based on data augmentation (Tanner and Wong, 1987) in the Bayesian approach, whereas the frequentist approach is usually based on the EM algorithm. MI is a Bayesian imputation method and the imputed estimating function is computed with respect to the posterior predictive distribution,

$$f(y_{\text{mis},i} \mid y_{\text{obs}}) = \int f(y_{\text{mis},i} \mid y_{\text{obs},i}; \theta)\pi(\theta \mid y_{\text{obs}})\, d\theta,$$

which is the average of the predictive distribution $f(y_{\text{mis},i} \mid y_{\text{obs},i}; \theta)$ over the posterior distribution of θ. On the other hand, in the frequentist approach, the conditional expectation in (2.3) is taken with respect to the prediction model (2.4) evaluated at $\theta = \hat{\theta}$, a consistent estimator of θ. For example, one can use the pseudo maximum likelihood estimator (MLE) $\hat{\theta}$ obtained by solving the pseudo mean score equation (Louis, 1982; Pfeffermann et al., 1998),

$$\bar{S}(\theta) = \sum_{i\in A} w_i E\{S(\theta; y_i) \mid y_{i,\text{obs}}; \theta\} = 0, \tag{2.5}$$

where $S(\theta; y_i) = \partial \log f(y_i; \theta)/\partial\theta$.

The Bayesian approach to imputation, especially in the context of MI, is well studied in the literature; however, to our best knowledge, there does not exist a comprehensive reference for FI, which is developed under the fully frequentist framework. In FI, the conditional expectation in (2.3) is computed by a weighted average of the imputed estimating functions

$$E\{U(\eta; y_i) \mid y_{\text{obs},i}\} \cong \sum_{j=1}^{M} w_{ij}^* U(\eta; y_{\text{obs},i}, y_{\text{mis},i}^{*(j)}), \tag{2.6}$$

where $\{y_{\text{mis},i}^{*(1)}, \ldots, y_{\text{mis},i}^{*(M)}\}$ are M imputed values for $y_{\text{mis},i}$, and $\{w_{i1}^*, \ldots, w_{iM}^*\}$ are the fractional weights that satisfy the conditions of $w_{ij}^* \geq 0$, $\sum_{j=1}^{M} w_{ij}^* = 1$ and

$$\sum_{i\in A} w_i \sum_{j=1}^{M} w_{ij}^* S(\hat{\theta}; y_{\text{obs},i}, y_{\text{mis},i}^{*(j)}) = 0.$$

TABLE 1
Comparison of two approaches of inference with missing data

	Bayesian	**Frequentist**
Model	Posterior distribution f(latent, θ | Obs.)	Prediction model f(latent | Obs., θ)
Learning algorithm	Data augmentation	EM algorithm
Prediction	Imputation(I)-step	Expectation(E)-step
Parameter update	Posterior(P)-step	Maximization(M)-step
Imputation	Multiple imputation	Fractional imputation
Variance estimation	Rubin's formula	Linearization or replication

Obs. indicates the observed data.

Once the FI data are constructed, the FI estimator of η is obtained by solving

$$\sum_{i\in A} w_i \sum_{j=1}^{M} w_{ij}^* U\bigl(\eta; y_{\mathrm{obs},i}, y_{\mathrm{mis},i}^{*(j)}\bigr) = 0. \tag{2.7}$$

Note that, by (2.6), (2.7) approximates the conditional estimating equation in (2.3).

In general, the FI method augments the original dataset as

$$\begin{aligned}\mathcal{S}_{\mathrm{FI}} = \bigl\{&\delta_i(w_i, y_i) + (1-\delta_i)\bigl(w_i w_{ij}^*, y_{ij}^*\bigr);\\ &j = 1, \ldots, M, i \in A\bigr\},\end{aligned} \tag{2.8}$$

where δ_i is the indicator of full response for y_i, and $y_{ij}^* = (y_{\mathrm{obs},i}, y_{\mathrm{mis},i}^{*(j)})$. That is, FI produces one single imputed dataset with size $\{np + n(1-p)M\}$, where n is the sample size, p is the proportion of full response, and M is the imputation size. In the fractionally imputed dataset, each unit with missing items is now represented by M copies and each copy is associated with an imputed value y_{ij}^* and a fractional weight w_{ij}^*. Since (2.6) holds for arbitrary U functions, the resulting estimator is approximately unbiased for a fairly large class of parameters, for example, domain means, percentages, quantiles, regression coefficients and correlations, which makes FI attractive for general-purpose estimation. Kim (2011) used the importance sampling technique to achieve (2.6) for general U functions, which will be presented in the next section.

3. PARAMETRIC FRACTIONAL IMPUTATION

Parametric Fractional Imputation (PFI), proposed by Kim (2011), features a parametric model approach to fractional imputation, and parameters in the imputation model are estimated by a computationally efficient EM algorithm.

To approximate the conditional estimating equation in (2.3) by PFI, for each missing value $y_{\mathrm{mis},i}$, we first generate M imputed values, denoted by $\{y_{\mathrm{mis},i}^{*(1)}, \ldots, y_{\mathrm{mis},i}^{*(M)}\}$ from a proposal distribution $h(y_{\mathrm{mis},i} \mid y_{\mathrm{obs},i})$. Section 6.2 provides some guidance for choosing a proposal distribution. Once the imputed values are generated from $h(\cdot)$, we compute

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

subject to $\sum_{j=1}^{M} w_{ij}^* = 1$, as the fractional weight assigned to $y_{ij}^* = (y_{\mathrm{obs},i}, y_{\mathrm{mis},i}^{*(j)})$, where $\hat{\theta}$ is the pseudo MLE of θ to be determined by the EM algorithm below. Since $\sum_{j=1}^{M} w_{ij}^* = 1$, the above fractional weight is the same as $w_{ij}^* = w_{ij}^*(\hat{\theta})$, where

$$w_{ij}^*(\theta) \propto \frac{f(y_{\mathrm{obs},i}, y_{\mathrm{mis},i}^{*(j)}; \theta)}{h(y_{\mathrm{mis},i}^{*(j)} \mid y_{\mathrm{obs},i})}, \tag{3.1}$$

which only requires knowledge of the joint distribution $f(y;\theta)$ and the proposal distribution h.

The pseudo MLE of θ can be computed by solving the imputed mean score equation

$$\sum_{i\in A} w_i \sum_{j=1}^{M} w_{ij}^*(\theta) S\bigl(\theta; y_{\mathrm{obs},i}, y_{\mathrm{mis},i}^{*(j)}\bigr) = 0, \tag{3.2}$$

where $w_{ij}^*(\theta)$ is defined in (3.1). To solve (3.2), we can either use the Newton method or the following EM algorithm:

I-*step*. For each missing value $y_{\mathrm{mis},i}$, M imputed values are generated from a proposal distribution $h(y_{\mathrm{mis},i} \mid y_{\mathrm{obs},i})$.

W-*step*. Using the current parameter value $\hat{\theta}_{(t)}$, compute the fractional weights as $w_{ij(t)}^* \propto f(y_{\mathrm{obs},i}, y_{\mathrm{mis},i}^{*(j)}; \hat{\theta}_{(t)}) / h(y_{\mathrm{mis},i}^{*(j)} \mid y_{\mathrm{obs},i})$, subject to $\sum_{j=1}^{M} w_{ij(t)}^* = 1$.

M-*step*. Update the parameter $\hat{\theta}_{(t+1)}$ by solving the imputed score equation,

$$\sum_{i\in A} w_i \sum_{j=1}^{M} w_{ij(t)}^* S(\theta; y_{ij}^*) = 0, \tag{3.3}$$

where $y_{ij}^* = (y_{\mathrm{obs},i}, y_{\mathrm{mis},i}^{*(j)})$.

C-*step*. Monitor the weight distribution using histograms or summary statistics. If there are extreme fractional weights that dominate other fractional weights, go to the I-step and modify the proposal distribution to a more plausible distribution, such as $f(y_{\mathrm{mis},i} \mid y_{\mathrm{obs},i}; \theta)$ evaluated at the current parameter value $\hat{\theta}_{(t)}$.

Iteration. Set $t = t + 1$ and go to the W-step. Stop if $\hat{\theta}_{(t+1)}$ meets the convergence criterion.

Here, the I-step is the imputation step, the W-step is the weighting step, and the M-step is the maximization step. The I- and W-steps can be combined to implement the E-step of the EM algorithm. Unlike the Monte Carlo EM (MCEM) method, imputed values are not changed for each EM iteration—only the fractional weights are changed. Thus, the FI method has computational advantages over the MCEM method. Convergence of the EM sequence of parameter estimates is

achieved because the imputed values are not changed unless the C-step has an effect. Kim (2011) showed that given the M imputed values, $y_{\text{mis},i}^{*(1)}, \ldots, y_{\text{mis},i}^{*(M)}$, the sequence of estimates $\{\hat{\theta}_{(0)}, \hat{\theta}_{(1)}, \ldots\}$ from the W- and M-steps converges to a stationary point $\hat{\theta}_M^*$ for fixed M. The stationary point $\hat{\theta}_M^*$ converges to the pseudo MLE of θ as $M \to \infty$. The resulting weight w_{ij}^* after convergence is the fractional weight assigned to $y_{ij}^* = (y_{\text{obs},i}, y_{\text{mis},i}^{*(j)})$. The C-step is used to assess the distribution of fractional weights. If several extremely large weights dominate other weights, it indicates that the proposal distribution is not well specified. A simple remedy is to update the imputation model with $\hat{\theta}_{(t)}$ and go to the I-step. Also, the C-step assists with convergence for finite M.

Once the fractionally imputed dataset is constructed, it can be used to estimate other parameters of interest. That is, we can use (2.7) to estimate η from the FI dataset.

We first consider a simple example to illustrate the idea of FI, which resembles regression imputation.

EXAMPLE 1. Suppose a probability sample consists of n units of $z_i = (x_i, y_i)$ with sampling weight w_i, where x_i is always observed and y_i is subject to missingness. Suppose the joint distribution in (2.1) is $f(x, y; \theta) = f(x) f(y \mid x; \theta)$. Under MAR, the MLE of θ can be obtained from the full respondent sample. After obtaining the MLE $\hat{\theta}$, M imputed values are generated for each missing y_i from $f(y_i \mid x_i; \hat{\theta})$. The imputed values $y_i^{*(1)}, \ldots, y_i^{*(M)}$ are assigned fractional weights $\omega_{ij}^*(\hat{\theta}) = 1/M$ since $h(y_i \mid x_i) = f(y_i \mid x_i; \hat{\theta})$. Then we can use (2.7) to estimate η from the FI dataset.

We now consider a bivariate missing data example to illustrate the use of the EM algorithm in FI.

EXAMPLE 2. Suppose a probability sample consists of n units of $z_i = (x_i, y_{1i}, y_{2i})$ with sampling weight w_i, where x_i is always observed and $y_i = (y_{1i}, y_{2i})$ is subject to missingness. Let A_{11}, A_{10}, A_{01} and A_{00} be the partitions of the sample based on the missing pattern, where the subscript $1/0$ in the ith position denotes that the ith y item is observed/missing, respectively. For example, A_{10} is the set of the sample with y_{1i} observed and y_{2i} missing.

The conditional expectation in (2.3) involves evaluating the conditional distribution of $y_{\text{mis},i}$ given the observed data x_i and $y_{\text{obs},i}$ for each missing pattern, which is then decomposed into

$$\begin{aligned}
&\sum_{i \in A} w_i E\{U(\eta; z_i) \mid x_i, y_{\text{obs},i}\} \\
&\quad = \sum_{i \in A_{11}} w_i U(\eta; x_i, y_{1i}, y_{2i}) \\
&\qquad + \sum_{i \in A_{00}} w_i E\{U(\eta; x_i, Y_{1i}, Y_{2i}) \mid x_i\} \\
&\qquad + \sum_{i \in A_{01}} w_i E\{U(\eta; x_i, Y_{1i}, y_{2i}) \mid x_i, y_{2i}\} \\
&\qquad + \sum_{i \in A_{10}} w_i E\{U(\eta; x_i, y_{1i}, Y_{2i}) \mid x_i, y_{1i}\}.
\end{aligned}$$

Suppose the joint distribution in (2.1) is

$$\begin{aligned}
&f(x, y_1, y_2; \theta) \\
&\quad = f_x(x; \theta_0) f_1(y_1 \mid x; \theta_1) f_2(y_2 \mid x, y_1; \theta_2).
\end{aligned} \tag{3.4}$$

From the full respondent sample in A_{11}, obtain $\hat{\theta}_{1(0)}$ and $\hat{\theta}_{2(0)}$, which are initial parameter estimates for θ_1 and θ_2.

In the I-step, for each missing value $y_{\text{mis},i}$, we generate M imputed values from $h(y_{\text{mis},i} \mid x_i, y_{\text{obs},i}) = f(y_{\text{mis},i} \mid x_i, y_{\text{obs},i}; \hat{\theta}_{(0)})$, where

$$\begin{aligned}
&f(y_{\text{mis},i} \mid x_i, y_{\text{obs},i}; \hat{\theta}_{(0)}) \\
&\quad = \begin{cases}
f_2(y_{2i} \mid x_i, y_{1i}; \hat{\theta}_{2(0)}), & \text{if } i \in A_{10}, \\
f(y_{1i} \mid x_i, y_{2i}; \hat{\theta}_{(0)}), & \text{if } i \in A_{01}, \\
f(y_{1i}, y_{2i} \mid x_i; \hat{\theta}_{(0)}), & \text{if } i \in A_{00}
\end{cases}
\end{aligned} \tag{3.5}$$

and

$$\begin{aligned}
&f(y_{1i} \mid x_i, y_{2i}; \hat{\theta}_{(0)}) \\
&\quad = \frac{f_1(y_{1i} \mid x_i; \hat{\theta}_{1(0)}) f_2(y_{2i} \mid x_i, y_{1i}; \hat{\theta}_{2(0)})}{\int f_1(y_{1i} \mid x_i; \hat{\theta}_{1(0)}) f_2(y_{2i} \mid x_i, y_{1i}; \hat{\theta}_{2(0)}) \, dy_{1i}}.
\end{aligned} \tag{3.6}$$

Note that the marginal distribution of x, $f_x(x; \theta_0)$, is not used in (3.6). Except for some special cases such as when both f_1 and f_2 are normal distributions, the conditional distribution in (3.6) is not in a known form. Thus, a computational tool such as Metropolis–Hastings (Hastings, 1970) or Sampling Importance Resampling (SIR; Rubin, 1987) is needed to generate samples from (3.6) for $i \in A_{01}$. For example, the SIR consists of the following steps:

1. Generate B (say $B = 100$) values, denoted by $y_{1i}^{*(1)}, \ldots, y_{1i}^{*(B)}$, from $f_1(y_{1i} \mid x_i; \hat{\theta}_{1(0)})$.

2. Among the B values obtained from Step 1, select one value with the selection probability proportional to $f_2(y_{2i} \mid x_i, y_{1i}^{*(k)}; \hat{\theta}_{2(0)})$, where $y_{1i}^{*(k)}$ is the kth value from Step 1 ($k = 1, \ldots, B$).
3. Repeat Step 1 and Step 2 independently M times to obtain M imputed values.

Once we obtain M imputed values of y_{1i} for $i \in A_{01}$, we can use

$$h(y_{1i} \mid x_i, y_{2i}) \propto f_1(y_{1i} \mid x_i; \hat{\theta}_{1(0)}) f_2(y_{2i} \mid x_i, y_{1i}; \hat{\theta}_{2(0)})$$

as the proposal density in (3.5). Since $\sum_{j=1}^{M} w_{ij}^* = 1$, we do not need to compute the normalizing constant in (3.6). For $i \in A_{10}$, M imputed values of y_{2i} are generated from $f_2(y_{2i} \mid x_i, y_{1i}; \hat{\theta}_{2(0)})$. For $i \in A_{00}$, M imputed values of y_{1i} can be generated from $f_1(y_{1i} \mid x_i; \hat{\theta}_{1(0)})$, and then M imputed values of y_{2i} can be generated from $f_2(y_{2i} \mid x_i, y_{1i}^*; \hat{\theta}_{2(0)})$.

In the W-step, the fractional weights are computed by

$$w_{ij(t)}^* \propto \frac{f_1(y_{1i}^{*(j)} \mid x_i; \hat{\theta}_{1(t)}) f_2(y_{2i}^{*(j)} \mid x_i, y_{1i}^{*(j)}; \hat{\theta}_{2(t)})}{h(y_{\mathrm{mis},i}^{*(j)} \mid x_i, y_{\mathrm{obs},i})}$$

with $\sum_{j=1}^{M} w_{ij(t)}^* = 1$, where $y_{1i}^{*(j)} = y_{1i}$ if y_{1i} is observed and $y_{2i}^{*(j)} = y_{2i}$ if y_{2i} is observed. Using the fractional weights, the parameter estimates are updated by the M-step in (3.3).

The above example covers a broad range of applications in the missing data literature, such as missing covariate problems, measurement error models, generalized linear mixed models, and so on. Yang and Kim (2016) considered regression analyses with missing covariates in survey data using FI, where in the current notation, $f(y_2 \mid x, y_1)$ is a regression model with y_2 and x fully observed and y_1 subject to missingness. In generalized linear mixed models, $f(y_2 \mid x, y_1)$ is a generalized linear mixed model where y_1 is the latent random effect. See Yang, Kim and Zhu (2013) for using FI to estimate parameters in generalized linear mixed models.

For variance estimation, note that the imputed estimator $\hat{\eta}_{\mathrm{FI}}$ obtained from the imputed estimating equation (2.7) depends on $\hat{\theta}$ obtained from (3.2). To reflect this dependence, we can write $\hat{\eta}_{\mathrm{FI}} = \hat{\eta}_{\mathrm{FI}}(\hat{\theta})$. To account for the sampling variability of $\hat{\theta}$ in the imputed estimator $\hat{\eta}_{\mathrm{FI}}$, either the linearization method or replication methods can be used. In the linearization method, the imputation model is needed in order to compute partial derivatives of the score functions. In some situations, disclosing the imputation model may not be desirable for confidentiality reasons. To avoid disclosing the imputation model, replication methods are often preferred (Rao and Shao, 1992). To implement replication variance estimation in FI, we first obtain the kth replicate pseudo MLE $\hat{\theta}^{[k]}$ of $\hat{\theta}$ by solving

$$\bar{S}^{*[k]}(\theta) \equiv \sum_{i \in A} w_i^{[k]} \sum_{j=1}^{M} w_{ij}^*(\theta) S(\theta; y_{ij}^*) = 0, \tag{3.7}$$

where $w_i^{[k]}$ is the kth replicate sampling weight and $w_{ij}^*(\theta)$ is defined in (3.1). To obtain $\hat{\theta}^{[k]}$ from (3.7), we can use either the EM algorithm as before or the one-step Newton method to ease the computational burden. For the one-step Newton method, we have

$$\hat{\theta}^{[k]} \cong \hat{\theta} - \left\{ \frac{\partial}{\partial \theta^T} \bar{S}^{*[k]}(\hat{\theta}) \right\}^{-1} \cdot \sum_{i \in A} w_i^{[k]} \sum_{j=1}^{M} w_{ij}^*(\hat{\theta}) S(\hat{\theta}; y_{ij}^*),$$

where

$$\frac{\partial}{\partial \theta^T} \bar{S}^{*[k]}(\theta) = \sum_{i \in A} w_i^{[k]} \sum_{j=1}^{M} w_{ij}^*(\theta) \dot{S}(\theta; y_{ij}^*) + \sum_{i \in A} w_i^{[k]} \sum_{j=1}^{M} w_{ij}^*(\theta) \left\{ S(\theta; y_{ij}^*) - \sum_{j=1}^{M} w_{ij}^*(\theta) S(\theta; y_{ij}^*) \right\}^{\otimes 2},$$

with $\dot{S}(\theta; y) = \partial S(\theta; y)/\partial \theta^T$ and $B^{\otimes 2} = BB^T$. Once $\hat{\theta}^{[k]}$ is obtained, we obtain the kth replicate $\hat{\eta}^{[k]}$ of $\hat{\eta}_{\mathrm{FI}}$ by solving

$$\sum_{i \in A} w_i^{[k]} \sum_{j=1}^{M} w_{ij}^{*[k]} U(\eta; y_{ij}^*) = 0$$

for η, where $w_{ij}^{*[k]} = w_{ij}^*(\hat{\theta}^{[k]})$. The replicates $\hat{\eta}_{\mathrm{FI}}^{[k]}$ are used to compute the estimator of the variance of $\hat{\eta}_{\mathrm{FI}}$,

$$\hat{V}_{\mathrm{rep}}(\hat{\eta}_{\mathrm{FI}}) = \sum_{k=1}^{L} c_k (\hat{\eta}_{\mathrm{FI}}^{[k]} - \hat{\eta}_{\mathrm{FI}})^2,$$

where L is the number of replicates, and c_k is the replication factor associated with replicate k.

4. NONPARAMETRIC FRACTIONAL IMPUTATION

4.1 Fractional Hot Deck Imputation

Hot deck imputation uses observed responses from the sample as imputed values. The unit with missing values is called the *recipient* and the unit providing the value for imputation is called the *donor*. Durrant (2009), Haziza (2009) and Andridge and Little (2010) provided comprehensive overviews of hot deck imputation in survey sampling. Hot deck imputation is very popular in household surveys.

Fractional hot deck imputation (FHDI) combines the ideas of FI and hot deck imputation. Kim and Fuller (2004), Fuller and Kim (2005), and Kim and Yang (2014) considered FHDI for univariate missing data. We now describe FHDI for multivariate missing data with an arbitrary missing pattern.

We first consider categorical data. Let $\underset{\sim}{z} = (z_1, \ldots, z_K)$ be the vector of study variables that take categorical values, and $\underset{\sim}{z}_i = (z_{i1}, \ldots, z_{iK})$ be the ith realization of $\underset{\sim}{z}$. Let δ_{ij} be the response indicator variable for z_{ij}. That is, $\delta_{ij} = 1$ if z_{ij} is observed and $\delta_{ij} = 0$ otherwise. Assume that the response mechanism is MAR. Based on $\delta_i = (\delta_{i1}, \ldots, \delta_{iK})$, $\underset{\sim}{z}_i$ can be decomposed into $(z_{\text{obs},i}, z_{\text{mis},i})$, which are the missing and observed parts of $\underset{\sim}{z}_i$, respectively. Let $D_i = \{z_{\text{mis},i}^{*(1)}, \ldots, z_{\text{mis},i}^{*(M_i)}\}$ be the set of all possible values of $z_{\text{mis},i}$, that is, $(z_{\text{obs},i}, z_{\text{mis},i}^{*(j)})$ is one of the actually observed values in the respondents. If all of M_i possible values in D_i are taken as the imputed values for $z_{\text{mis},i}$, the fractional weight assigned to the jth imputed value $z_{\text{mis},i}^{*(j)}$ is

$$w_{ij}^* = \frac{\pi(z_{\text{obs},i}, z_{\text{mis},i}^{*(j)})}{\sum_{k\in D_i} \pi(z_{\text{obs},i}, z_{\text{mis},i}^{*(k)})}, \tag{4.1}$$

where $\pi(\underset{\sim}{z})$ is the joint probability of $\underset{\sim}{z}$. Empirically, the joint probability can be approximated by

$$\pi(\underset{\sim}{z}) = \frac{\sum_{i\in A} w_i \sum_{j\in D_i} w_{ij}^* I\{(z_{\text{obs},i}, z_{\text{mis},i}^{*(j)}) = \underset{\sim}{z}\}}{\sum_{i\in A} w_i}. \tag{4.2}$$

The EM algorithm by weighting (Ibrahim, 1990) can be used to compute (4.1) and (4.2), starting with the initial values of fractional weights $w_{ij(0)}^* = M_i^{-1}$. Equations (4.1) and (4.2) correspond to the E-step and M-step of the EM algorithm, respectively. The M-step (4.2) can be changed if we assume a parametric model for the joint probability $\pi(\underset{\sim}{z})$. For example, if the joint probability is a parsimonious multinomial distribution $\pi(\underset{\sim}{z}; \alpha)$, then the M-step replaces (4.2) with solving the imputed score equation of α to update the estimate of α.

We now consider continuous data. Let $\underset{\sim}{y} = (y_1, \ldots, y_K)$ be the vector of study variables that take continuous values, and $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iK})$ be the ith realization of $\underset{\sim}{y}$. We can first discretize each continuous variable by dividing its range into a small finite number of segments. Let z_{ik} denote the discrete version of y_{ik}. Note that z_{ik} is observed only if y_{ik} is observed. The support of $\underset{\sim}{z}$, denoted by $\{\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_G\}$, is the same as the sample support of $\underset{\sim}{z}$ from the full respondents and specifies donor cells. The joint probability of $\underset{\sim}{z}$, denoted by $\pi(\underset{\sim}{z}_g)$, for $g = 1, \ldots, G$, can be obtained by the EM algorithm for categorical missing data as described above.

As in the categorical case, let $D_i = \{z_{\text{mis},i}^{*(1)}, \ldots, z_{\text{mis},i}^{*(M_i)}\}$ be the set of all possible values of $z_{\text{mis},i}$. Using a finite mixture model, a nonparametric approximation of $f(y_{\text{mis},i} \mid y_{\text{obs},i})$ is

$$f(y_{\text{mis},i} \mid y_{\text{obs},i}) \approx \sum_{j=1}^{M_i} P\big(\underset{\sim}{z} = \underset{\sim}{z}_i^{*(j)} \mid y_{\text{obs},i}\big) f\big(y_{\text{mis},i} \mid \underset{\sim}{z}_i^{*(j)}\big), \tag{4.3}$$

where each $\underset{\sim}{z}_i^{*(j)} = (z_{\text{obs},i}, z_{\text{mis},i}^{*(j)})$ defines an imputation cell. The approximation in (4.3) is based on the assumption that

$$f(y_{\text{mis}} \mid y_{\text{obs}}, \underset{\sim}{z}) \cong f(y_{\text{mis}} \mid \underset{\sim}{z}), \tag{4.4}$$

which means (approximate) conditional independence between y_{mis} and y_{obs} given $\underset{\sim}{z}$. Thus, we assume that the covariance structure between items are captured by the discrete approximation and the within-cell errors can be safely assumed to be independent. Once the imputation cells are formed to satisfy (4.4), we select m_g imputed values for $y_{\text{mis},i}$, denoted by $\underset{\sim}{y}_i^{*(j)} = (y_{\text{obs},i}, y_{\text{mis},i}^{*(j)})$, for $j = 1, \ldots, m_g$, randomly from the full respondents in the same cell, with the selection probability proportional to the sampling weights. The final fractional weight assigned to $\underset{\sim}{y}_i^{*(j)}$ is $w_{ij}^* = P(\underset{\sim}{z}_i^{*(j)} \mid y_{\text{obs},i}) m_g^{-1}$.

This FHDI procedure resembles a two-phase stratified sampling (Rao, 1973, Kim, Navarro and Fuller, 2006), where forming imputation cells corresponds

to stratification (Phase one) and conducting hot deck imputation corresponds to stratified sampling (Phase two).

If we select all possible donors in the same cell, the resulting FI estimator is fully efficient in the sense that it does not introduce additional randomness due to hot deck imputation. Such fractional hot deck imputation is called fully efficient fractional imputation (FEFI). FEFI is implemented at Proc Surveyimpute in SAS (SAS Institute Inc, 2015).

4.2 Nonparametric Fractional Imputation Using Kernels

In real-data applications, nonparametric methods are preferred if less is known about the true underlying data distribution. Hot deck imputation makes less or no parametric distributional assumptions and, therefore, is more robust than fully parametric methods. In this section, we discuss another way of calculating the fractional weights that links the FI estimator to some well-known nonparametric estimators, such as the Nadaraya–Watson kernel regression estimator (Nadaraya, 1964).

For simplicity, suppose we have bivariate data (x_i, y_i) where x_i is completely observed and y_i is subject to missingness. Assume the missing data mechanism is MAR. Let δ_i be the response indicator that takes value one if y_i is observed and zero otherwise. We are interested in estimating η, which is defined through $E\{U(\eta; X, Y)\} = 0$. Let $A_R = \{i \in A; \delta_i = 1\}$ be the index set of respondents, and $r = |A_R|$ be the size of A_R. To calculate the conditional estimating equation (2.3) nonparametrically, we use the following FI algorithm: for each unit i with $\delta_i = 0$, we take r values from A_R as imputed values of y_i, denoted by $y_i^{*(1)}, \ldots, y_i^{*(r)}$, and compute the kernel-based fractional weight $w_{ij}^* = K_h(x_i - x_i^{*(j)}) / \sum_{k \in A_R} K_h(x_i - x_i^{*(k)})$, where $K_h(\cdot)$ is the kernel function with bandwidth h and $x_i^{*(j)}$ is the covariate associated with $y_i^{*(j)}$. The nonparametric fractional weight measures the degree of similarity of y_i and $y_i^{*(j)}$ based on the distance between x_i and $x_i^{*(j)}$. The resulting FI estimating equation can be written as

$$\sum_{i \in A} w_i \left\{ \delta_i U(\eta; x_i, y_i) + (1 - \delta_i) \sum_{j \in A_R} w_{ij}^* U\big(\eta; x_i, y_i^{*(j)}\big) \right\} = 0. \tag{4.5}$$

The FI estimator uses $\hat{U}(\eta; x_i) = \sum_{j \in A_R} w_{ij}^* U(\eta; x_i, y_i^{*(j)})$ to approximate $E\{U(\eta; x_i, y_i) \mid x_i\}$ nonparametrically. For fixed η, $\hat{U}(\eta; x_i)$ is often called the *Nadaraya–Watson kernel regression estimator* of $E\{U(\eta; x_i, y_i) \mid x_i\}$ in the nonparametric estimation literature. Note that this FI estimator does not rely on any parametric model assumptions and so is nonparametric; however, it is not assumption free because it makes an implicit assumption of the continuity of $E\{U(\eta; x, y) \mid x_i\}$ through the choice of kernels to define the "similarity" (Nadaraya, 1964). Notably, while the convergence of $\hat{U}(\eta; x_i)$ to $E\{U(\eta; x_i, y_i) \mid x_i\}$ does not achieve the order of $O_p(1/\sqrt{n})$, the solution $\hat{\eta}_{\mathrm{FI}}$ to (4.5) satisfies $\hat{\eta}_{\mathrm{FI}} - \eta = O_p(1/\sqrt{n})$ under some regularity conditions, which was proved by Wang and Chen (2009) in the classical setup of independent and identically distributed observations.

Such kernel-based nonparametric fractional imputation is directly applicable to complex survey sampling. More developments are expected by coupling FI with other nonparametric methods such as those using nearest neighbor imputation (Chen and Shao, 2001; Beaumont and Bocci, 2009; Kitamura, Tripathi and Ahn, 2004; Kim, Fuller and Bell, 2011) or predictive mean matching (Vink et al., 2014).

5. SYNTHETIC DATA IMPUTATION

Synthetic imputation is a technique of creating imputed values for items not observed in the current survey by incorporating information from other surveys. For example, suppose that there are two independent surveys, called Survey 1 and Survey 2, and we observe x_i from Survey 1 and observe (x_i, y_i) from Survey 2. In this case, we may want to create synthetic values of y_i in Survey 1 by incorporating information from Survey 2, so that inference about y can be made even in Survey 1. Synthetic imputation is particularly useful when Survey 1 is a large scale survey and item y is very expensive to measure. Schenker and Raghunathan (2007) reported several applications of combining information from multiple surveys. In one application, they discussed synthetic imputation that combined information from two surveys conducted by the National Center for Health Statistics to improve on analyses of self-reported data on health conditions: in one survey, both self-reported health measurements and clinical measurements from physical examinations were available, and in the much larger survey, only self-reported health measurements were available.

The setup of two independent samples with common items can also be called non-nested two-phase sampling. Analyzing data from two-phase sampling can be treated as a missing data problem, where the missingness is planned and the response probability is known. In two-phase sampling, suppose we observe x_i in the first-phase sample and observe (x_i, y_i) in the second-phase sample, where the second-phase sample is not necessarily nested within the first-phase sample. Let A_1 and $\{w_{i1}; i \in A_1\}$ be the sets of indices and sampling weights for the first-phase sample, respectively. Let A_2 and $\{w_{i2}; i \in A_2\}$ be the corresponding sets for the second-phase sample. Assume a working model $m(x_i; \theta)$ for $E(y \mid x_i)$. For estimation of the population total of y, the two-phase regression estimator can be written as

$$\hat{Y}_{\mathrm{tp}} = \sum_{i \in A_1} w_{i1} m(x_i; \hat{\theta}) + \sum_{i \in A_2} w_{i2}\{y_i - m(x_i; \hat{\theta})\}, \tag{5.1}$$

where $\hat{\theta}$ is estimated from the second-phase sample. The two-phase regression estimator is efficient if the working model is well specified. The first term of (5.1) is called the projection estimator. Kim and Rao (2012) discussed asymptotic properties of the projection estimator under nonnested two-phase sampling. Note that if the second term of (5.1) is equal to zero, the two-phase regression estimator is equivalent to the projection estimator. Asymptotic properties of the two-phase estimator and variance estimation have been discussed in Kim, Navarro and Fuller (2006), and Kim and Yu (2011a).

Creating an imputed dataset for the first-phase sample, often called mass imputation, is one method for incorporating the second-phase information into the first-phase sample. Fuller (2003) investigated mass imputation in the context of two-phase sampling. In a large scale survey, it is a common practice to produce estimates for domains. Legg and Fuller (2009) discussed the possibility of using imputation to obtain improved estimators for domains.

The FI procedure can be used to obtain the two-phase regression estimator in (5.1) and, at the same time, improve domain estimation. Note that the two-phase regression estimator (5.1) can be written as

$$\hat{Y}_{\mathrm{FEFI}} = \sum_{i \in A_1} \sum_{j \in A_2} w_{i1} w_{ij}^* y_i^{*(j)}, \tag{5.2}$$

where $y_i^{*(j)} = \hat{y}_i + \hat{e}_j$, $\hat{y}_i = m(x_i; \hat{\theta})$, $\hat{e}_j = y_j - \hat{y}_j$, $w_{ij}^* = w_{j2}/(\sum_{k \in A_2} w_{k2})$, and we assume $\sum_{i \in A_1} w_{i1} = \sum_{i \in A_2} w_{i2}$. The expression (5.2) implies that we impute all the units in the first-phase sample, including the units that also belong to the second-phase sample. The estimator (5.2) is computed using an augmented dataset of $n_1 \times n_2$ records, where n_1 and n_2 are the sizes of A_1 and A_2, respectively, and the (i, j)th record has an (imputed) observation $y_i^{*(j)} = \hat{y}_i + \hat{e}_j$ with weight $w_{i1} w_{ij}^*$. That is, for each unit $i \in A_1$, we impute n_2 values of $y_i^{*(j)}$ with fractional weight w_{ij}^*. The method in (5.2) imputes all the units in A_2 and is called fully efficient fractional imputation (FEFI), according to Fuller and Kim (2005). The FEFI estimator is algebraically equivalent to the two-phase regression estimator of the population total of y, and can also provide consistent estimators for other parameters such as population quantiles.

If it is desirable to limit the number of imputations to a small value m ($m < n_2$), FI using regression weighting in Fuller and Kim (2005) can be adopted. We first select m values of $y_i^{*(j)}$, denoted by $y_i^{**(1)}, \ldots, y_i^{**(m)}$, from the set of n_2 imputed values $\{y_i^{*(j)}; j \in A_2\}$ using an efficient sampling method. The fractional weights $\tilde{w}_{ij}^*$ assigned to the selected values $y_i^{**(j)}$ are determined so that

$$\sum_{j=1}^{m} \tilde{w}_{ij}^* (1, y_i^{**(j)}) = \sum_{j \in A_2} w_{ij}^* (1, y_i^{*(j)}) \tag{5.3}$$

holds for each $i \in A_1$. The fractional weights satisfying (5.3) can be computed using the regression weighting method or the empirical likelihood method; see Section 6.1 for details. The resulting FI data $y_i^{**(j)}$ with weights $w_{i1} \tilde{w}_{ij}^*$ are constructed with $n_1 \times m$ records, which integrate available information from two phases. Replication variance estimation with FI, similar to Fuller and Kim (2005), can be developed. See Section 8.7 of Kim and Shao (2014).

6. FRACTIONAL IMPUTATION VARIANTS

6.1 The Choice of M and Calibration Fractional Imputation

The choice of the imputation size M is a matter of tradeoff between statistical efficiency and computational efficiency: small M may lead to large variability in the Monte Carlo approximation; whereas large M may increase computational cost. The magnitude of the imputation error is usually $O(1/\sqrt{M})$, which can be reduced for large M. Thus, if computational power allows, the larger M, the better.

In survey practice, it is not desirable to release a large imputed dataset to the public. To reduce the size of the final dataset, a subset of initial imputated values can be selected. In this case, FI can be developed in three stages. The first stage, called *Fully Efficient Fractional Imputation* (FEFI), computes the pseudo MLE of parameters in the superpopulation model with a sufficiently large imputation size M, say $M = 1000$. The second stage, called the *Sampling* stage, selects a small number m (say, $m = 10$) imputed values from the initial set of M imputed values. The third stage, called *Calibration Weighting*, constructs the final fractional weights for the m selected imputed values to satisfy calibration constraints. This procedure can be called *Calibration FI*.

The FEFI step is described in Section 5. Here, we describe the last two stages in detail. In the Sampling stage, a subset of imputed values is selected into reduce the final imputation size. For each i, we have M initial imputed values $y_{ij}^* = (y_{\text{obs},i}, y_{\text{mis},i}^{*(j)})$ with fractional weights w_{ij}^*. We treat $\underset{\sim}{y_i}^* = \{y_{ij}^*; j = 1, \ldots, M\}$ as a weighted finite population with weights $\{w_{ij}^*; j = 1, \ldots, M\}$, and use an unequal probability sampling method such as probability-proportion-to-size (PPS) sampling without replacement to select a sample of size m, say $m = 10$, from $\underset{\sim}{y_i}^*$ using w_{ij}^* as the selection probability. Let $\tilde{y}_{i1}^*, \ldots, \tilde{y}_{im}^*$ be the m values sampled from $\underset{\sim}{y_i}^*$.

According to the above selection rule, the fractional weights for the sampled m imputed values are given by $\tilde{w}_{ij0}^* = m^{-1}$. However, this set of fractional weights may not necessarily satisfy the imputed score equation,

$$\sum_{i \in A} w_i \sum_{j=1}^{m} \tilde{w}_{ij}^* S(\hat{\theta}; \tilde{y}_{ij}^*) = 0, \tag{6.1}$$

where $\hat{\theta}$ is the pseudo MLE of θ obtained at the FEFI stage. It is desirable for the solution to the imputed score equation with small m to be equal to the pseudo MLE of θ, which specifies the calibration constraint. At the Calibration Weighting stage, the initial set of weights is modified to satisfy the constraints (6.1) and $\sum_{j=1}^{m} \tilde{w}_{ij}^* = 1$, which can be achieved by regression weighting. The regression fractional weights are constructed by

$$\tilde{w}_{ij}^* = \tilde{w}_{ij0}^* + \tilde{w}_{ij0}^* \Delta (S_{ij}^* - \bar{S}_i^*), \tag{6.2}$$

where $S_{ij}^* = S(\hat{\theta}; y_{ij}^*)$, $\bar{S}_i^* = \sum_{j=1}^{m} \tilde{w}_{ij0}^* S_{ij}^*$, and $\Delta = -\{\sum_{i \in A} w_i \bar{S}_i^*\}^T \{\sum_{i \in A} w_i \sum_{j=1}^{m} \tilde{w}_{ij0}^* (S_{ij}^* - \bar{S}_i^*)^{\otimes 2}\}^{-1}$. Note that some of the fractional weights computed by (6.2) may be negative. To avoid negative weights, alternative algorithms other than regression weighting can be used. For example, the fractional weights of the form

$$\tilde{w}_{ij}^* = \frac{\tilde{w}_{ij0}^* \exp(\Delta S_{ij}^*)}{\sum_{k=1}^{m} \tilde{w}_{ik0}^* \exp(\Delta S_{ik}^*)}$$

are approximately equal to the regression fractional weights in (6.2) and are always positive.

6.2 The Choice of the Proposal Distribution

PFI generates imputed values from the *proposal distribution* h. The choice of the proposal distribution is somewhat arbitrary. However, a well-specified proposal distribution may improve the finite-sample performance of the imputation estimator. In what follows, we discuss a number of ways to specify the proposal distribution and assess the quality of specification.

For a planned parameter, for example, the population mean of y, Kim (2011) showed the optimal proposal distribution that makes the Monte Carlo approximation variance of $\bar{y}_I^* = \sum_{j=1}^{M} w_{ij}^* y_{ij}^*$ as small as possible, is

$$\begin{aligned} h^*(y_{\text{mis},i} \mid y_{\text{obs},i}) &= f(y_{\text{mis},i} \mid y_{\text{obs},i}; \hat{\theta}) \\ &\quad \cdot \frac{|y_i - E\{y_i \mid y_{\text{obs},i}; \hat{\theta}\}|}{E\{|y_i - E\{y_i \mid y_{\text{obs},i}; \hat{\theta}\}| \mid y_{\text{obs},i}; \hat{\theta}\}}, \end{aligned}$$

where $\hat{\theta}$ is the MLE of θ. For general-purpose estimation, the parameter of interest is often unplanned at the time of imputation. According to Fay (1992), $h(y_{\text{mis},i} \mid y_{\text{obs},i}) = f(y_{\text{mis},i} \mid y_{\text{obs},i}; \hat{\theta})$ is a reasonable choice in terms of statistical efficiency. For importance sampling, since we do not know $\hat{\theta}$ at the outset of the EM algorithm, we can use $h(y_{\text{mis},i} \mid y_{\text{obs},i}) = \int f(y_{\text{mis},i} \mid y_{\text{obs},i}; \theta) \pi(\theta)\, d\theta$, where $\pi(\theta)$ is a prior distribution for θ.

We now discuss a special choice of the proposal distribution h, based on the realized values of the variables having missing values, which is akin to hot deck imputation. For simplicity, assume that the scalar variable y_i is subject to missingness, δ_i is the response indicator of y_i, and x_i is completely observed in the sample. For each missing value y_i, $A_R = \{j \in A; \delta_j = 1\}$ forms a donor pool. For this choice of imputed values, the proposal distribution $h(y_j)$ is $f(y_j \mid \delta_j = 1)$. In calculating the fractional weights, we approximate $f(y_j \mid \delta_j = 1)$ by its empirical distribution $N_R^{-1} \sum_{k \in A} w_k \delta_k f(y_j \mid x_k)/N_R$, where $N_R = \sum_{i \in A} w_k \delta_k$. The EM algorithm takes the following steps:

I-*step*. For each missing value y_i, take all values in A_R as donors.

W-*step*. With current parameter estimates of θ, $\hat{\theta}_{(t)}$, compute the fractional weights by

$$w^*_{ij(t)} \propto \sum_{k\in A} w_k \delta_k f(y_j \mid x_k), \tag{6.3}$$

subject to $\sum_{j\in A} \delta_j w^*_{ij(t)} = 1$.

M-*step*. Update the parameter $\hat{\theta}_{(t+1)}$ by solving the following imputed score equation:

$$\sum_{i\in A} w_i \delta_i S(\theta; x_i, y_i) + \sum_{i\in A} w_i(1-\delta_i) \sum_{j\in A_R} w^*_{ij(t)} S(\theta; x_i, y_j) = 0.$$

Iteration. Set $t = t+1$ and go to the W-step. Stop if $\hat{\theta}_{(t+1)}$ meets the convergence criterion.

Once the FI dataset is created, the FI estimator of the population mean $\bar{Y}$ is

$$\hat{\bar{Y}}_{\mathrm{FI}} = \frac{1}{N}\left\{\sum_{i\in A} w_i \delta_i y_i + \sum_{i\in A} w_i(1-\delta_i) \sum_{j\in A_R} w^*_{ij} y_j\right\}.$$

Kim and Yang (2014) showed that the resulting estimator gains robustness over PFI, due to the special choice of the proposal distribution. It is less sensitive to departures from the assumed parametric model.

6.3 Doubly Robust Fractional Imputation

Suppose we have bivariate data (x_i, y_i) where x_i is completely observed, y_i is subject to missingness and the missing data mechanism is MAR. We assume an outcome regression (OR) model $E(y_i \mid x_i) = m(x_i; \theta_0)$ and a response propensity (RP) model $P(\delta_i = 1 \mid x_i, y_i) = P(\delta_i = 1 \mid x_i) = \pi(x_i; \phi_0)$. Let $A_R = \{i; \delta_i = 1\}$ be the set of respondents, where δ_i is the response indicator of y_i. We are interested in estimating the population total $\eta = \sum_{i=1}^N y_i$. Notice that we do not need both the OR and RP models to construct consistent estimators of η. For example, $\hat{\eta}_1 = \sum_{i\in A} w_i m(x_i; \hat{\theta})$, with $\hat{\theta}$ being a consistent estimator of θ_0, is consistent for η under the OR model and $\hat{\eta}_2 = \sum_{i\in A_R} w_i y_i / \pi(x_i; \hat{\phi})$, with $\hat{\phi}$ being a consistent estimator of ϕ_0, is consistent for η under the RP model.

An estimator of η is doubly robust if it is consistent if either the OR model or the RP model is correctly specified, but not necessarily both. This property guards the estimator from bias due to model misspecification. The DR estimators have been extensively studied in the literature; see for example, Robins, Rotnitzky and Zhao (1994), Bang and Robins (2005), Tan (2006), Kang and Schafer (2007), Cao, Tsiatis and Davidian (2009), and Kim and Haziza (2014). We now discuss a FI estimator that has the double robustness feature.

Let $m(x; \theta)$ be fitted to the respondent set A_R, leading to a consistent estimator of θ, $\hat{\theta}$, under the OR model. For each unit $j \in A_R$, we have the residual $\hat{e}_j = y_j - m(x_j; \hat{\theta})$. For each missing value y_i, let the donor pool be A_R, and $y^*_{ij} = \hat{y}_i + \hat{e}_j$ be the jth imputed value, where $\hat{y}_i = m(x_i; \hat{\theta})$ and $\hat{e}_j$ is contributed from the donor $j \in A_R$. Each donor probably does not contribute equally. We use inverse probability weighting to determine the fractional weight associated with the imputed value. We fit $\pi(x; \phi)$ such that $\sum_{i\in A_R} w_i \{\pi(x_j; \hat{\phi})\}^{-1} = \sum_{i\in A} w_i$, so that each unit $j \in A_R$ represents $1/\pi(x_j; \hat{\phi})$ copies of the sample. Then the fractional weight w^*_{ij} associated with the jth imputed value y^*_{ij} is proportional to $\{1/\pi(x_j; \hat{\phi}) - 1\}$ over the donor pool A_R (minus one because y_j itself counts one), that is,

$$w^*_{ij} = \frac{w_j\{1/\pi(x_j; \hat{\phi}) - 1\}}{\sum_{k\in A} w_k \delta_k \{1/\pi(x_k; \hat{\phi}) - 1\}}. \tag{6.4}$$

The FI estimator is given by

$$\hat{\eta}_{\mathrm{FI}} = \sum_{i\in A} w_i \left[\delta_i y_i + (1-\delta_i)\left\{\sum_{j\in A} \delta_j w^*_{ij} y^*_{ij}\right\}\right]. \tag{6.5}$$

The FI estimator $\hat{\eta}_{\mathrm{FI}}$ in (6.5) is doubly robust. First, notice that $\hat{\eta}_{\mathrm{FI}}$ is algebraically equal to

$$\hat{\eta}_{\mathrm{FI}} = \sum_{i\in A} w_i \left[m(x_i; \hat{\theta}) + \frac{\delta_i}{\pi(x_i; \hat{\phi})}\{y_i - m(x_i; \hat{\theta})\}\right]. \tag{6.6}$$

Let $\hat{\eta}_n = \sum_{i\in A} w_i y_i$ be the full sample estimator of η, then

$$\hat{\eta}_{\mathrm{FI}} - \hat{\eta}_n = \sum_{i\in A} w_i \left\{\frac{\delta_i}{\pi(x_i; \hat{\phi})} - 1\right\}\{y_i - m(x_i; \hat{\theta})\}.$$

This is an asymptotically unbiased estimator of zero if either the OR model or the RP model is correctly specified, but not necessarily both. Kim and Haziza (2014) discussed efficient estimation of (θ, ϕ) and doubly robust variance estimation in survey sampling.

7. COMPARISON WITH MULTIPLE IMPUTATION

Multiple imputation (MI) has been proposed as a general tool for imputation and features simplified variance estimation. It is therefore of interest to compare

the behavior of MI and FI. In this section, we compare MI and FI in terms of statistical efficiency and their ability to handle informative sampling.

7.1 Statistical Efficiency

For the purpose of illustration, we consider a simple setting where the complete data z are randomly drawn from a population whose density is $f(z;\theta)$ with $\theta \in \mathbb{R}^d$. MI creates M complete datasets by imputing the missing data z_{mis} from the posterior predictive distribution given the observed data z_{obs}, $f(z_{\text{mis}} \mid z_{\text{obs}}) = \int f(z_{\text{mis}} \mid z_{\text{obs}};\theta)\pi(\theta \mid z_{\text{obs}})\,d\theta$, where $\pi(\theta \mid z_{\text{obs}})$ is the posterior distribution of θ. The MI estimator of θ is

$$\hat{\theta}_{\text{MI}} = M^{-1}\sum_{k=1}^{M}\hat{\theta}^{(k)},$$

where $\hat{\theta}^{(k)}$ is the MLE applied to the kth imputed dataset. Rubin's variance formula is

$$\hat{V}_{\text{MI}}(\hat{\theta}_{\text{MI}}) = W_M + \bigl(1+M^{-1}\bigr)B_M,$$

where $W_M = M^{-1}\sum_{k=1}^{M}\hat{V}^{(k)}$, $B_M = (M-1)^{-1}\times \sum_{k=1}^{M}(\hat{\theta}^{(k)}-\hat{\theta}_{\text{MI}})^2$, and $\hat{V}^{(k)}$ is the variance estimator of $\hat{\theta}$ under complete response applied to the kth imputed dataset.

Bayesian MI is a simulation-based method, and thus introduce additional variability in generating parameters from the posterior distribution. This explains why the asymptotic variance of the MI estimator, given by Wang and Robins (1998),

$$\begin{aligned} V_{\text{MI}} &= \mathcal{I}_{\text{obs}}^{-1} + M^{-1}\mathcal{I}_{\text{com}}^{-1}\mathcal{I}_{\text{mis}}\mathcal{I}_{\text{com}}^{-1} \\ &\quad + M^{-1}J^T\mathcal{I}_{\text{obs}}^{-1}J, \end{aligned} \tag{7.1}$$

is strictly larger than the asymptotic variance of the FI estimator,

$$V_{\text{FI}} = \mathcal{I}_{\text{obs}}^{-1} + M^{-1}\mathcal{I}_{\text{com}}^{-1}\mathcal{I}_{\text{mis}}\mathcal{I}_{\text{com}}^{-1}, \tag{7.2}$$

where $\mathcal{I}_{\text{com}} = E\{S(\theta)^{\otimes 2}\}$, $\mathcal{I}_{\text{obs}} = E\{S_{\text{obs}}(\theta)^{\otimes 2}\}$, $\mathcal{I}_{\text{mis}} = \mathcal{I}_{\text{com}} - \mathcal{I}_{\text{obs}}$, $S(\theta) = S(Z;\theta) = \partial \log f(Z;\theta)/\partial\theta$ is the score function of the complete-data likelihood, $S_{\text{obs}}(\theta) = E\{S(\theta) \mid Z_{\text{obs}}\}$ is the score function of the observed-data likelihood, and $J = \mathcal{I}_{\text{mis}}\mathcal{I}_{\text{com}}^{-1}$ is the fraction of missing information matrix (Rubin, 1987, Chapter 4). The difference between (7.1) and (7.2) can be sizable for a small M. Furthermore, for a large M, although the MI estimator is efficient, the inference is inefficient since Rubin's estimator of the variance of the MI estimator is only weakly unbiased, that is, $\hat{V}_{\text{MI}}(\hat{\theta}_{\text{MI}})$ converges in distribution instead of converges in probability to V_{MI}. This leads to much broader confidence intervals and less powerful tests than produced via a consistent variance estimator (Nielsen, 2003). On the other hand, the replication variance estimator for FI, discussed in Section 3, is consistent for V_{FI}.

7.2 Imputation Under Informative Sampling

Under informative sampling, the MAR assumption is subtle. We assume that the response mechanism is MAR at the population level, now referred to as population missing at random (PMAR), to be distinguished from the concept of sample missing at random (SMAR). For simplicity, assume y is a scalar variable which is subject to missingness, δ is the response indicator of y, x is a vector of covariates which is always observed, and I is the sample inclusion indicator. PMAR means that $y \perp \delta \mid x$, that is, MAR holds at the population level, $f(y \mid x) = f(y \mid x, \delta)$. On the other hand, SMAR is defined as $y \perp \delta \mid (x, I=1)$, that is, MAR holds at the sample level, $f(y \mid x, I=1) = f(y \mid x, I=1, \delta)$. The two assumptions are not testable empirically. The plausibility of these assumptions should be judged by subject matter experts. Often, PMAR is regarded to be more realistic than SMAR because an individual's decision about whether or not to respond to a survey more likely depends on his or her own characteristics than on being actually being selected; for example, a person may never respond to any telephone or internet survey as a general principle.

With an noninformative sampling design, we have $P(I=1 \mid x, y) = P(I=1 \mid x)$, under which PMAR implies SMAR. With informative sampling designs, however, PMAR does not necessarily imply SMAR. In such cases, using an imputation model fitted to the sample data for generating imputed values can result in biased estimators.

FI does not require that SMAR holds in addition to PMAR. Under PMAR, we have $f(y \mid x, \delta=0) = f(y \mid x)$. Let $f(y \mid x;\theta)$ be a parametric model for imputed values $f(y \mid x)$. The parameter θ can be consistently estimated by solving (2.5), even under informative sampling. Since FI generates the imputations from $f(y \mid x;\hat{\theta})$, with a consistent estimator $\hat{\theta}$, the resulting FI estimator is approximately unbiased (Berg, Kim and Skinner, 2016), whereas, MI tends to problematic under informative sampling if SMAR does not hold. To address this, a number of researchers suggest using an augmented model by adding sampling weights (or some function of sampling weights) into the imputation model to achieve SMAR, claiming that

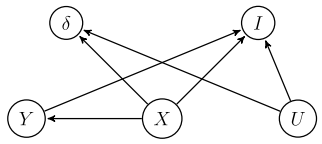

FIG. 1. *A directed acyclic graph (DAG) for a setup where PMAR holds but SMAR does not hold. Variable U is latent in the sense that it is never observed.*

the resultant MI point estimator is approximately unbiased (Rubin, 1996; Schenker et al., 2006). However, as pointed out by Berg, Kim and Skinner (2016), this approach is not always successful. For example, as presented in Figure 1, Y is conditionally independent of δ given X. However, Y is not conditionally independent of δ given X and I when there exists a latent variable U that affects both δ and I, because I is a collider in the pathway from Y to δ. In this case, augmenting X by including sampling weights does not achieve SMAR and leaves inference from the MI estimator questionable.

8. SIMULATION STUDY

In this section, we investigated the performance of FI compared to MI by a limited simulation study using an artificial finite population generated from real survey data. The pseudo finite population was generated from a single month of the US Census Bureau's Monthly Retail Trade Survey (MRTS). Each month, the MRTS surveys a sample of about 12,000 retail businesses with paid employees to collect data on sales and inventories. The MRTS is an economic indicator survey whose monthly estimates are inputs to the Gross Domestic Product estimates. The MRTS sample design is a typical one for business surveys, employing one-stage stratified sampling with stratification based on major industry classification, further substratified by the estimated annual sales. The sample design specifies higher sampling rates in strata with larger units than in strata with smaller units. More detail about MRTS and the simulated data can be found in Mulry, Oliver and Kaputa (2014).

The original population file contains 19,601 retail businesses stratified into 16 strata, with a strata identifier (h), sales (y, 10^4 US dollars), and inventory values (x, 10^4 US dollars). For simulation purposes, we focused on the first 5 strata as a finite population, consisting of 7260 retail businesses. Figure 2 shows the scatter plot of the sales and inventory data by strata on a log scale. We assumed the following superpopulation model:

$$\log(y_{hi}) = \beta_{0h} + \beta_{1h}\log(x_{hi}) + \varepsilon_{hi}, \tag{8.1}$$

where β_{0h} and β_{1h} are strata-specific parameters with h being the strata identifier, and $\varepsilon_{hi} \sim N(0, \sigma_h^2)$. Figure 3 shows the residual plot and the normal Q–Q plot for the fitted model (8.1) to assess the adequacy of model (8.1). From the residual plot, the constant variance assumption of ε_{hi} within strata appears to be reasonable. From the normal Q–Q plot, the normality assumption of ε_{hi} holds approximately.

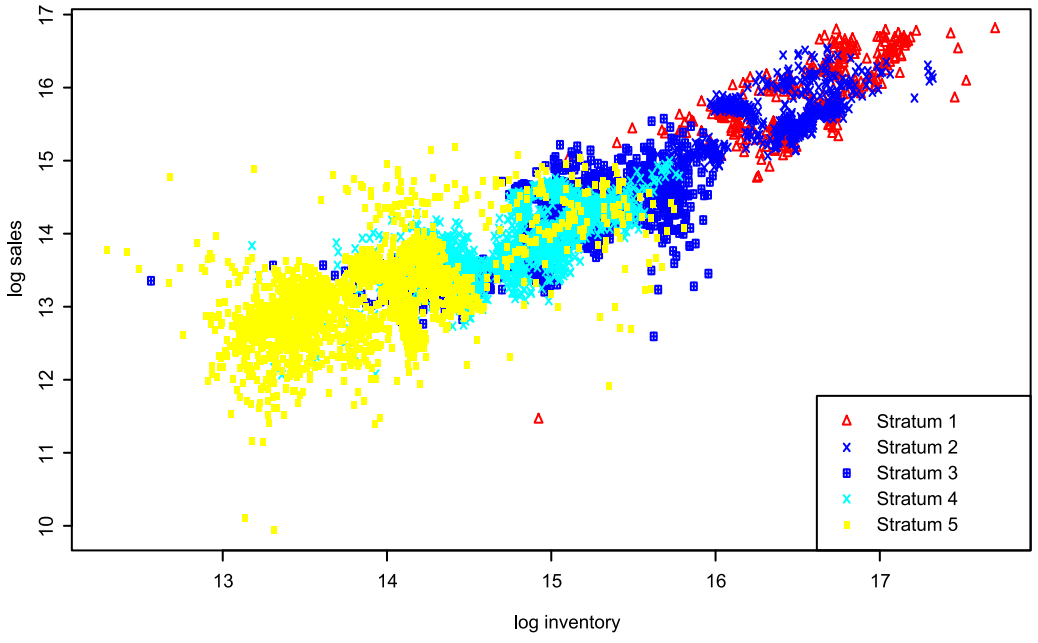

FIG. 2. *Scatter plot of log sales and log inventory data by strata.*

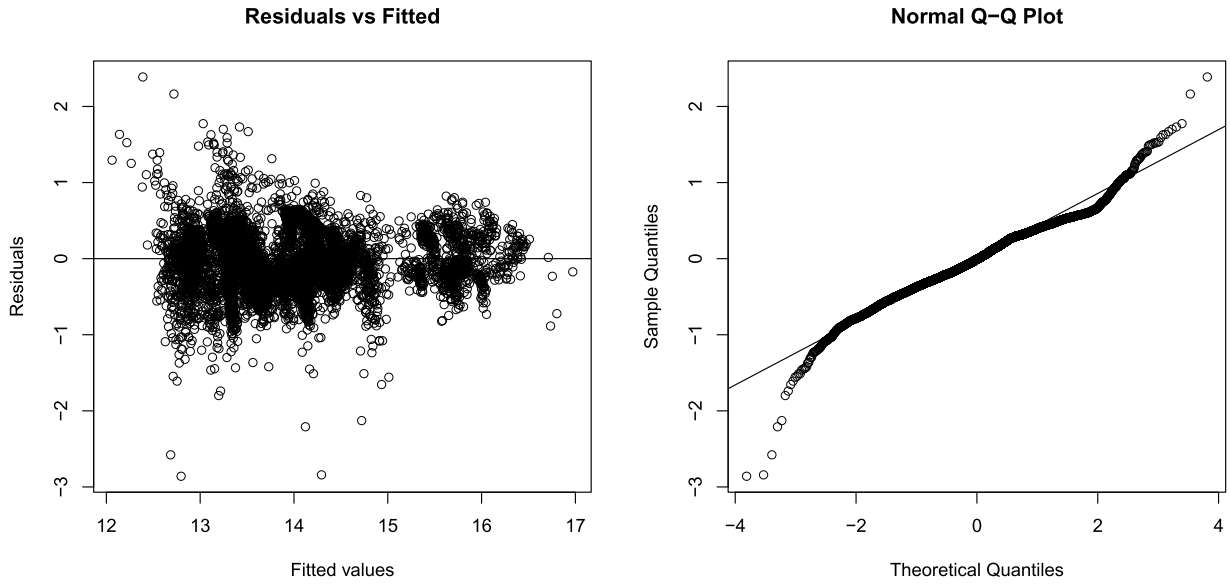

FIG. 3. *Diagnostic plots for the regression of* $\log(y)$ *on* $\log(x)$ *and stratum indicators*: *the residual plot* (*left*) *and the normal Q–Q plot* (*right*).

To create missing data, we considered univariate missingness where only y has missing values. We generated the response indicator δ for y according to

$$\delta \sim \text{Bernoulli}(\pi), \quad \pi = 1/[1 + \exp\{4 - 0.3\log(x)\}].$$

Under this model, the missing mechanism is MAR and the response rate is about 0.6.

To generate samples, we considered stratified sampling with simple random sampling within strata (STSRS). Table 2 shows strata sizes N_h, sample sizes n_h, and sampling weights in each stratum. The sampling weights range from 13 to 46. The selection of samples is repeated 2000 times.

The parameters of interest are the stratum means of y, $\eta_h = \mu_h$ for $1 \le h \le 5$, and the population mean of y, $\eta_6 = \mu$. The true parameter values are $\eta_1 = 92.25$, $\eta_2 = 67.90$, $\eta_3 = 18.24$, $\eta_4 = 13.01$, $\eta_5 = 5.92$ and $\eta_6 = 20.40$. The estimation methods included (i) Full, the full sample estimator, which is used as a benchmark for comparison, (ii) MI, the multiple imputation estimator with imputation size $M = 100$ and (iii) PFI, the parametric fractional imputation estimator with imputation size $M = 100$, where the model parameters are estimated by the pseudo MLE solving the imputed score equation (3.2).

TABLE 2
The sample allocation in stratified simple random sampling

Strata S_h	S_1	S_2	S_3	S_4	S_5
Strata size N_h	352	566	1963	2181	2198
Sample size n_h	28	32	46	46	48
Sampling weight	13	18	43	47	46

For both MI and PFI, we used the log normal regression model in (8.1) as the imputation model. Because the sampling design is stratified random sampling and the imputation model includes the stratum indicators, the sampling design becomes noninformative. We first imputed $\log(y)$ from the posterior predictive distribution for (8.1), given the observed data, and then transformed the imputations to the original scale of y. In each imputed dataset, we applied the design-unbiased full-sample point estimators and variance estimators for the STSRS design.

For PFI, the proposal distribution in the importance sampling step is the imputation distribution evaluated at initial parameter values estimated from the available data. For estimating model parameters, we obtained the pseudo MLEs by solving the score equation (2.5). After imputation, η was estimated by solving (2.7) by choosing U to be the corresponding estimating function. We used the delete-1 Jackknife replication method for variance estimation,

$$\hat{V}_R(\hat{\eta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i \in S_h} \big(\hat{\eta}^{[i]} - \hat{\eta}\big)^2,$$

where $\hat{\eta}^{[i]}$ is computed by omitting unit $i \in S_h$ and modifying the weights so that the sampling weight w_{hj}

TABLE 3

Numerical Results of Point Estimation (Mean and Var), Relative Bias (R.B.) of Variance Estimation, Mean Width and Coverage of 95% Confidence Intervals (C.I.s) under Stratified Simple Random Sampling over 2000 Samples. The estimation methods include (i) *FULL: the full sample estimator,* (ii) *MI: the multiple imputation estimator with imputation size* $M = 100$, (iii) *PFI, the parametric fractional imputation estimator with imputation size* $M = 100$. *The parameters are* $\eta_1 =$ *Stratum 1 mean,* $\eta_2 =$ *Stratum 2 mean,* $\eta_3 =$ *Stratum 3 mean,* $\eta_4 =$ *Stratum 4 mean,* $\eta_5 =$ *Stratum 5 mean,* $\eta_6 =$ *Population mean*

	Mean			Var			R.B. (%)			Mean width of C.I.s			Coverage		
	FULL	MI	PFI	FULL	MI	PFI	FULL	MI	PFI	FULL	MI	PFI	FULL	MI	PFI
η_1	92.46	93.95	92.85	76.46	119.18	120.67	6.08	48.06	7.81	18.01	26.57	22.81	0.951	0.964	0.952
η_2	67.72	68.40	67.76	40.05	60.91	59.53	6.55	30.53	3.26	13.07	17.83	15.68	0.943	0.954	0.946
η_3	18.30	18.45	18.28	2.12	3.32	3.29	−3.06	23.05	−1.63	2.86	4.04	3.60	0.944	0.961	0.948
η_4	13.03	13.12	13.00	1.02	1.77	1.76	0.51	23.02	−4.28	2.03	2.95	2.60	0.946	0.962	0.943
η_5	5.92	5.98	5.91	0.22	0.46	0.46	1.84	16.96	−4.40	0.94	1.47	1.32	0.953	0.963	0.947
η_6	20.42	20.63	20.42	0.70	1.11	1.10	−3.36	32.75	−3.97	1.65	2.42	2.06	0.952	0.983	0.953

is replaced by $n_h w_{hj}/(n_h - 1)$ for all $j \in S_h$ and the sampling weights remain the same for all other units.

Table 3 shows the numerical results. The means and variances were calculated as the Monte Carlo means and Monte Carlo variances of the point estimates across 2000 simulated datasets. The relative bias of the variance estimator was calculated as $\{(\text{Ve} - \text{Var})/\text{Var}\} \times 100\%$, where Ve is the Monte Carlo mean of variance estimates and Var is Monte Carlo variance of point estimates. In addition, 95% confidence intervals were calculated. We obtained the Monte Carlo mean widths and coverages of 95% confidence intervals. The three estimators are essentially unbiased for point estimation, which is expected since the full sample estimator is design-consistent, and for MI and PFI, the imputation model is correctly specified. Our primary interest lies in comparison of the performance of variance estimation. As shown in Table 3, the mean width of confidence intervals based on MI is larger than that of FI. Rubin's estimator of the variance of the MI estimator is biased upward with relative biases 48.06%, 30.53%, 23.05%, 23.02% and 16.96% for $\hat{\eta}_{j,\text{MI}}$, $1 \le j \le 5$ and 32.75% for $\hat{\eta}_{6,\text{MI}}$, respectively. Because of this variance overestimation, the empirical coverage of 95% confidence interval reaches 98.3% for the population mean. Rubin's variance estimator requires a self-efficient complete-sample estimator (Meng, 1994), even when the congeniality condition is satisfied, that is, the imputation model is correctly specified as in our simulation study. An estimator $\hat{\eta}(\cdot)$ is self-efficient if it never decreases the variance when it is applied to any subset of the data compared to the complete data, that is,

$$V\{\hat{\eta}(Y_{\text{sub}})\} \ge V\{\hat{\eta}(Y_{\text{com}})\}, \tag{8.2}$$

where Y_{sub} is any subset of the data and Y_{com} is the complete data. Otherwise, Rubin's variance estimator is biased (Meng and Romero, 2003). Self-efficiency holds for the maximum likelihood estimator of η. Under the log normal distribution and MAR, the design-unbiased estimators are the method of moments estimators, which were used as complete-sample estimators. The (design-unbiased) Horvitz–Thompson estimators are not self-efficient in the log-normal model, which explains the bias in Rubin's variance estimator. In contrast, the FI variance estimator is essentially unbiased and the empirical coverage of 95% confidence intervals is close to the nominal coverage level.

9. CONCLUDING REMARKS

Both multiple imputation (MI) and fractional imputation (FI) can be used to create complete datasets for general-purpose estimation from sample survey data (subject to missingness). Rubin's MI variance formula is simple and easy to use, but its validity requires special conditions called congeniality and self-efficiency that can be restrictive in practice. In contrast, FI does not require the self-efficiency condition for consistent variance estimation. When the sampling design is informative, MI can use an augmented model to make the sampling design noninformative. However, incorporating all design information into the model is not always possible (Reiter, Raghunathan and Kinney, 2006). When this happens, valid inference of MI is not easy and is sometimes impossible (Berg, Kim and Skinner, 2016). In contrast, FI can handle informative sampling more easily as it incorporates sampling weights into estimation instead of modeling.

So far we have presented FI under MAR, but other response mechanisms can be considered. Parametric FI can be adapted to situations where the missing values are suspected to be missing not at random (MNAR) (Kim and Kim, 2012; Yang, Kim and Zhu, 2013). A semiparametric FI using the exponential tilting model of Kim and Yu (2011b) is also promising, which is under development. Also, FI can be used to approximate the observed log likelihood (Yang and Kim, 2016a), which can be directly applied to model selection or model comparison with missing data, such as using the Akaike Information Criterion (Akaike, 1998) or the Bayesian Information Criterion (Schwarz, 1978). Further investigation of this topic will be worthwhile.

We conclude with the hope that continuing efforts will be made into developing statistical methods and corresponding computational programs for FI (an R software package is in progress), so as to make these methods accessible to a broader audience. Proc SurveyImpute in SAS (version 14.1) contains some options for fractional imputation for categorical data.

ACKNOWLEDGMENTS

We are grateful to Emily Berg, Katherine J. Thompson, the Associate Editor and two referees for their valuable comments that helped to improve this paper. The research of the second author was partially supported by a grant from US National Science Foundation and also by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University.

REFERENCES

AKAIKE, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* 199–213. Springer, Berlin.

ANDRIDGE, R. R. and LITTLE, R. J. (2010). A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* **78** 40–64.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BEAUMONT, J.-F. and BOCCI, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canad. J. Statist.* **37** 400–416. MR2547206

BEAUMONT, J.-F., HAZIZA, D. and BOCCI, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statist. Sinica* **21** 515–537. MR2829845

BERG, E., KIM, J. K. and SKINNER, C. (2016). Imputation under informative sampling. *Surv. Methodol.* To appear.

BINDER, D. A. and PATAK, Z. (1994). Use of estimating functions for estimation from complex surveys. *J. Amer. Statist. Assoc.* **89** 1035–1043. MR1294748

BINDER, D. A. and SUN, W. (1996). Frequency valid multiple imputation for surveys with a complex design. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* 281–286. Amer. Statist. Assoc., Alexandria, VA.

CAO, W., TSIATIS, A. A. and DAVIDIAN, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96** 723–734. MR2538768

CHAUVET, G., DEVILLE, J.-C. and HAZIZA, D. (2011). On balanced random imputation in surveys. *Biometrika* **98** 459–471. MR2806441

CHEN, J. and SHAO, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *J. Amer. Statist. Assoc.* **96** 260–269. MR1952736

DURRANT, G. B. (2005). Imputation methods for handling item-nonresponse in the social sciences: a methodological review. ESRC National Centre for Research Methods and Southampton Stat. Sci. Research Institute. NCRM Methods Review Papers NCRM/002.

DURRANT, G. B. (2009). Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates. *International Journal of Social Research Methodology* **12** 293–304.

DURRANT, G. B. and SKINNER, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Surv. Methodol.* **32** 25–36.

FAY, R. E. (1992). When are inferences from multiple imputation valid? In *Proceedings of the Survey Research Methods Section of the American Statistical Association* **81** 227–332. Amer. Statist. Assoc., Alexandria, VA.

FAY, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *J. Amer. Statist. Assoc.* **91** 490–498.

FULLER, W. A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data* (*Southampton*, 1999) (R. L. Chambers and C. J. Skinner, eds.) 307–322. Wiley, Chichester. MR1978858

FULLER, W. A. and KIM, J. K. (2005). Hot deck imputation for the response model. *Surv. Methodol.* **31** 139–149.

GODAMBE, V. P. and THOMPSON, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *Int. Stat. Rev.* **54** 127–138. MR0962931

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57** 97–109.

HAZIZA, D. (2009). Imputation and inference in the presence of missing data. In *Sample Surveys*: *Design*, *Methods and Applications* (C. R. Rao and D. Pfeffermann, eds.). *Handbook of Statist.* **29** 215–246. Elsevier, Amsterdam. MR2654640

IBRAHIM, J. G. (1990). Incomplete data in generalized linear models. *J. Amer. Statist. Assoc.* **85** 765–769.

KALTON, G. and KISH, L. (1984). Some efficient random imputation methods. *Comm. Statist. Theory Methods* **13** 1919–1939.

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458

KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98** 119–132. MR2804214

KIM, J. K. and FULLER, W. (2004). Fractional hot deck imputation. *Biometrika* **91** 559–578. MR2090622

KIM, J. K., FULLER, W. A. and BELL, W. R. (2011). Variance estimation for nearest neighbor imputation for US census long form data. *Ann. Appl. Stat.* **5** 824–842. MR2840177

KIM, J. K. and HAZIZA, D. (2014). Doubly robust inference with missing data in survey sampling. *Statist. Sinica* **24** 375–394. MR3183689

KIM, J. K. and HONG, M. (2012). Imputation for statistical inference with coarse data. *Canad. J. Statist.* **40** 604–618. MR2968401

KIM, J. Y. and KIM, J. K. (2012). Parametric fractional imputation for nonignorable missing data. *J. Korean Statist. Soc.* **41** 291–303. MR3255335

KIM, J. K., NAVARRO, A. and FULLER, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *J. Amer. Statist. Assoc.* **101** 312–320. MR2268048

KIM, J. K. and RAO, J. N. K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika* **99** 85–100. MR2899665

KIM, J. K. and SHAO, J. (2014). *Statistical Methods for Handling Incomplete Data*. Chapman & Hall, Raton, FL. MR3307946

KIM, J. K. and YANG, S. (2014). Fractional hot deck imputation for robust inference under item nonresponse in survey sampling. *Surv. Methodol.* **40** 211–230.

KIM, J. K. and YU, C. L. (2011a). Replication variance estimation under two-phase sampling. *Surv. Methodol.* **37** 67–74.

KIM, J. K. and YU, C. L. (2011b). A semiparametric estimation of mean functionals with nonignorable missing data. *J. Amer. Statist. Assoc.* **106** 157–165. MR2816710

KIM, J. K., BRICK, J. M., FULLER, W. A. and KALTON, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 509–521. MR2278338

KITAMURA, Y., TRIPATHI, G. and AHN, H. (2004b). Empirical likelihood-based inference in conditional moment restriction models. *Econometrika* **72** 1667–1714.

KOTT, P. (1995). A paradox of multiple imputation. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* 384–389.

LEGG, J. C. and FULLER, W. A. (2009). Two-phase sampling. In *Sample Surveys*: *Design*, *Methods and Applications*. *Handbook of Statist.* **29** 55–70. Elsevier, Amsterdam. MR2654633

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233. MR0676213

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9** 538–558.

MENG, X.-L. and ROMERO, M. (2003). Discussion: Efficiency and self-efficiency with multiple imputation inference. *Int. Stat. Rev.* **71** 607–618.

MULRY, M. H., OLIVER, B. E. and KAPUTA, S. J. (2014). Detecting and treating verified influential values in a monthly retail trade survey. *J. Off. Stat.* **30** 721–747.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

NIELSEN, S. F. (2003). Proper and improper multiple imputation. *Int. Stat. Rev.* **71** 593–607.

PFEFFERMANN, D., SKINNER, C. J., HOLMES, D. J., GOLDSTEIN, H. and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 23–56. MR1625668

RAO, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika* **60** 125–133. MR0331576

RAO, J. N. K. and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79** 811–822. MR1209480

RAO, J. N. K., YUNG, W. and HIDIROGLOU, M. A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhya*, *Ser. A* **64** 364–378. MR1981764

REITER, J. P., RAGHUNATHAN, T. E. and KINNEY, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Surv. Methodol.* **32** 143.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.

SAS INSTITUTE INC (2015). SAS/STAT 14.1 User's Guide—the SURVEYIMPUTE Procedure. SAS Institue Inc., Cary, NC.

SCHAFER, J. L. (1997). Imputation of missing covariates under a multivariate linear mixed model. Unpublished technical report.

SCHENKER, N. and RAGHUNATHAN, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Stat. Med.* **26** 1802–1811. MR2359193

SCHENKER, N., RAGHUNATHAN, T. E., CHIU, P.-L., MAKUC, D. M., ZHANG, G. and COHEN, A. J. (2006). Multiple imputation of missing income data in the National Health interview survey. *J. Amer. Statist. Assoc.* **101** 924–933. MR2324093

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

VINK, G., FRANK, L. E., PANNEKOEK, J. and VAN BUUREN, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Stat. Neerl.* **68** 61–90. MR3168318

WANG, D. and CHEN, S. X. (2009). Empirical likelihood for estimating equations with missing values. *Ann. Statist.* **37** 490–517. MR2488360

WANG, N. and ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948. MR1666715

WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

YANG, S. and KIM, J. K. (2016). A semiparametric inference to regression analysis with missing covariates in survey data. *Statist. Sinica*. To appear.

YANG, S. and KIM, J. K. (2016a). Likelihood-based inference with missing data under missing-at-random. *Scand. J. Stat.* **43** 436–454.

YANG, S. and KIM, J. K. (2016b). A note on multiple imputation for method of moments estimation. *Biometrika* **103** 244–251. MR3465836

YANG, S., KIM, J.-K. and ZHU, Z. (2013). Parametric fractional imputation for mixed models with nonignorable missing data. *Stat. Interface* **6** 339–347. MR3105224