Variance estimation by multivariate imputation methods in complex survey designs

Jong-Min Kim\textsuperscript{a}, Kee-Jae Lee\textsuperscript{b} and Wonkuk Kim\textsuperscript{c,}\textsuperscript{*}
\textsuperscript{a}Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, USA
\textsuperscript{b}Department of Information Statistics, Korea National Open University, Seoul, Korea
\textsuperscript{c}Department of Applied Statistics, Chung-Ang University, Seoul, Korea

Abstract. In this paper, we consider variance estimation of the sample mean when the missing data have been imputed with multivariate imputation methods. Modern multivariate imputation methods to missing data are complicated and computationally expensive. These multivariate imputation methods do not require the normality assumption to impute the missing values. Under this assumption free condition, we compare the performance of variance estimation of six modern multivariate imputation methods including copula imputation, random forest imputation, principal component analysis imputation, and k-nearest neighbors imputation methods in complex sampling designs such as stratified sampling, cluster sampling and resampling approach to variance estimation by jackknife and bootstrap methods in stratified sampling. We conducted simulation studies using National Health and Nutrition Survey data considering 5\% and 15\% missing completely at random (MCAR) rates. Based on our 500 times resampling simulation study of the mean squares errors of the sample mean in complex survey designs, the percent relative efficiency (RE(\%)) of the random forest (RF) imputation method appears to outperform other imputation methods overall when the data has high skewness at the 5\% missing rate and when the data has high excessive kurtosis at the 15\% missing rate whereas the principal component analysis (PCA) imputation method appears to outperform other imputation methods when the data has high skewness at the 5\% and 15\% missing rates. Especially, the RE(\%) of the multivariate imputation methods appears to be efficient in the cluster sampling design when the data has high skewness or excessive kurtosis at the 15\% missing rate.

Keywords: Missing at random (MAR), copula imputation, jackknife, bootstrap

1. Introduction

Estimation of population variance is of significant importance in the theory of estimation in presence of nonresponse. Efficient variance estimation under auxiliary information has been widely discussed by various authors such as Singh (2003), Arnab and Singh (2006), Kim et al. (2007) and Singh et al. (2015).

A common significant issue in survey research is nonresponse at both the unit and the item level. Unit nonresponse (weight adjustment) refers to the complete absence of an interview from an eligible sample member whereas item nonresponse (missing data) refers to the absence of answers to a particular item or items in the interview after the eligible sampled member agrees to participate in the survey. Survey statisticians have dealt with unit and item nonresponse as two different problems with separate impacts on data quality, different statistical analysis results which lead to the misleading conclusions, and different underlying causes (Kim & Anderson, 2004; Groves, 2004; Groves et al. 2004; Groves et al. 2000).
Missing data make the statistical analysis of the study more complicated and less efficient. Missing values may reduce the representativeness of the samples, induce bias in estimation of parameters, and lead to a loss of power in hypothesis tests. The statistical method of replacing missing item with plausible values from incomplete observations from the collected or available data is called imputation which enables construction of standard programs based on some probability sampling model, for substituting missing data with a point estimate. Missing data imputation is a statistical technique that replaces missing values with statistical predicted values so that imputation method has potential to reduce bias and improve precision to a significant extent. Singh (2003) and Kim and Shao (2013) have reviewed various imputation techniques.

Rubin (1976) introduced the term missing completely at random (MCAR) to describe data where the complete cases are a random sample of the originally identified set of cases, that is, the probability of a case having a missing value for a variable does not depend on either the known values or the missing data. Missing at random (MAR) is another missing mechanism in which the probability of a case having a missing value for a variable may depend on the known values but not on the value of the missing data itself. When the probability of a case having a missing value for a variable depends on the value of that variable, it is called missing not at random (MNAR). Heitjan and Basu (1996) have distinguished the meaning of MAR and missing completely at random (MCAR) in a very nice way. The clarification of these two methods includes the idea that the parameter spaces for models involving the data and the response mechanism are distinct.

In the last two decades an increasing interest has been observed in proposing many statistical methods for handling missing values. Survey researchers have commonly used simple methods such as complete case analysis (listwise deletion), available case analysis (pairwise deletion), or single-value imputation and multiple imputation methods which produce multiple imputed datasets to take into account the uncertainty of the imputation. Some of modern imputation approaches implement multivariate statistical techniques and machine learning algorithms such as principal component analysis, singular value decomposition, copulas, k-nearest neighbors, or random forest, etc. Most of published articles in this area deal with the development of new imputation methods. There are some recent work to compare and evaluate a variety of existing imputation techniques in order to provide practical guidelines for researchers to make an appropriate choice of imputation methods in different situations (Schmitt et al. 2015). Recently, Kowarik and Templ (2016) summarized a variety of imputation techniques available in the R packages, SPSS and SAS.

Though there are some comparison work on imputation methods (Schmitt et al., 2015; Celton et al. 2010; Saunders et al., 2006; De Brevern et al. 2004; Brock et al. 2008), none of them have evaluated performance of the imputation methods by connecting variance estimation in complex survey designs.

Imputation is commonly employed for item nonresponse in sample surveys and then compute the variance estimates using standard formulas. In this paper, we report the relative efficiency of variance estimations of the sample mean of some modern and traditional multivariate imputation techniques under complex survey designs such as stratified sampling, cluster sampling, bootstrap sampling, and jackknife sampling. To achieve that goal, taking the National Health and Nutrition Survey data, datasets were randomly simulated at rates 5% and 15% of missingness, imposing the MCAR mechanism in the original database.

The remainder of this paper is organized as follows: Section 2 describes the modern multivariate imputation methods for missing data and the variance estimation in complex survey designs. Section 3 shows the dataset description and empirical data simulation study. Finally, conclusions are presented in Section 4.

2. Method

We begin this section by summarizing multivariate imputation methods considered in this paper.

2.1. Modern multivariate imputation methods for missing data

1. Copula imputation: Käärik (2006) and E. Käärik and M. Käärik (2009) used Gaussian copulas to impute correlated incomplete data with repeated measurements. Since then, diverse copula-based imputation methods have been proposed. Di Lascio et al. (2015) emphasized the advantage of the copula based imputation is inherited
from those of copula functions that make it possible to model parsimoniously complex multivariate dependence structures allowing to remove the assumption of normality for the data and to fit each margin through the most appropriate model. Zeisberger (2014) also proposed vine copula regression imputation, vine copula fitting imputation and vine copula expectation imputation based on vine dependence structure. Di Lascio and Giannerini (2016) developed an Coint R package for an imputation method based on copulas. Suppose there are $p$ continuous variables $X_1, \ldots, X_p$ with marginal distributions $F_1, \ldots, F_p$, respectively. We let $U_j = F_j(X_j)$ be the integral transform of the $j$th variable. The joint density can be written as
\[
f(x_1, \ldots, x_p) = c(F_1(x_1), \ldots, F_p(x_p)) \prod_{j=1}^p f_j(x_j)
\]
where $c(\cdot)$ is a copula density. The conditional copula density
\[
c(u_j|u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_p)
\]
is defined by Bayes’ rule. The conditional density of $X_j$ given other variables may be calculated by
\[
f(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p) = c(u_j|u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_p)f_j(x_j)
\]
It can be easily extended to the case of multivariate missing variables.
\[
f(x_j, x_{j'}|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{j'-1}, x_{j'+1}, \ldots, x_p) = c(u_j, u_{j'}|u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_{j'-1}, u_{j'+1}, \ldots, u_p)f_j(x_j)f_{j'}(x_{j'})
\]
We can impute missing observations by means of the Hit or Miss Monte Carlo method.

2. Random forest imputation (RF): Random forest (Breiman, 2001) is one of machine learning approaches. Random forest imputation is an iterative nonparametric imputation method based on random forest. For each variable, we fit a random forest on the observed data and then predict the missing values. This algorithm continues to repeat these two steps until a stopping criterion is met or the pre-specified maximum of iterations is reached. R package missForest (Stekhoven & Bühlmann, 2012) is running iteratively, continuously updating the imputed matrix variable-wise, and is assessing its performance between iterations. This assessment is done by considering the difference(s) between the previous imputation result and the new imputation result. As soon as this difference (in case of one type of variable) or differences (in case of mixed-type of variables) increase, the algorithm stops. Random forest offers advantages in terms of dealing with mixed-type data, it is relatively robust to outliers and noise, and does not assume normality, linearity, or homoscedasticity (Hill, 2012; Rieger et al. 2010).

3. Predictive mean matching (PMM): PMM (Little, 1988; Rubin, 1986) is another way to do multiple imputation for missing data, particularly for imputing continuous variables that are not normally distributed. Suppose there is a single variable $x$ that has some cases with missing data, and a set of variables $z$ (with no missing data) that are used to impute $x$.
   a. For cases with no missing data, estimate a linear regression of $x$ on $z$, producing a set of coefficients $\beta$.
   b. Make a random draw from the “posterior predictive distribution” of $\beta$, producing a new set of coefficients $\beta^*$. Typically this would be a random draw from a multivariate normal distribution with mean $\beta$ and the estimated covariance matrix of $\beta$ (with an additional random draw for the residual variance). This step is necessary to produce sufficient variability in the imputed values, and is common to all “proper” methods for multiple imputation.
   c. Using $\beta^*$, generate predicted values for $x$ for all cases, both those with data missing on $x$ and those with data present.
   d. For each case with missing $x$, identify a set of cases with observed $x$ whose predicted values are close to the predicted value for the case with missing data.
   e. From among those close cases, randomly choose one and assign its observed value to substitute for the missing value.
   f. Repeat steps b through e for each completed data set.
4. Principal component analysis imputation (PCA): Alternative common approach dealing with missing values in PCA consists in ignoring the missing values by minimizing the least squares criterion over all non missing entries. This can be achieved by the introduction of a weighted matrix $W$ in the criterion, with $w_{ik} = 0$ if $x_{ik}$ is missing and $w_{ik} = 1$, otherwise:

$$
C = \sum_{i=1}^{I} \sum_{k=1}^{K} w_{ik} \left( x_{ik} - m_k - \sum_{s=1}^{S} F_{is} U_{ks} \right)^2
$$

There is no explicit solution to minimize the criterion Eq. (4) when missing values are present and it is necessary to use iterative algorithms. Many algorithms are available and (Gabriel & Zamir, 1979) proposed an iterative PCA algorithm that an imputation of the missing values is achieved during the estimation process. Their algorithm consists of the following steps:

a. Missing values are replaced by initial values $X^0$ such as the mean of each variable.

b. PCA is performed on the completed data set to estimate the parameters and obtain $(\hat{M}^\ell, \hat{F}^\ell, \hat{U}^\ell)$ while the dimension $S$ is fixed.

c. Missing values are imputed with the fitted values $\hat{X}^\ell = M^\ell + \hat{F}^\ell (\hat{U}^\ell)^t$; the new imputed dataset is $X^\ell = W \ast X + (1 - W) \ast \hat{X}^\ell$: observed values are the same and missing values are replaced by the fitted one.

d. Steps b and c of imputation are repeated until convergence.

5. k-nearest neighbors (kNN) imputation: k-nearest neighbors algorithm can be used to estimate and substitute missing data. The kNN can predict both discrete variables by majority voting and continuous variables by averaging. For a continuous variable,

$$
\hat{x} = \frac{1}{|N|} \sum_{x_i \in N(x)} x_i
$$

This kNN approach does not produce explicit prediction models. Thus, the k-nearest neighbor can be easily adapted to work with any attribute as class, by just modifying which attributes will be considered in the distance metric. Also, this approach can easily treat examples with multiple missing values. The main drawback of this approach is that whenever the k-nearest neighbor looks for the most similar instances, the algorithm searches through all the data set. This limitation can be very critical for the analysis of large databases.

6. Singular value decomposition imputation (SVD): SVD can be used in a simple way to impute data to missing values (Krzanowski, 1988). The method is easy to compute and a description of the steps for one missing value $x_{ij}$ in $X$ followed:

a. Omit the $i$th case (row) from $X$ and calculate the SVD of the remaining $(n - 1) \times p$ data matrix, denoted by $X^{-i} = \bar{U} \bar{V}$ with $\bar{U} = \{ \bar{u}_1, ..., \bar{u}_p \}$ and $\bar{V} = \{ \bar{v}_1, ..., \bar{v}_n \}$, where $\bar{U}$ and $\bar{V}$ are orthonormal matrices (i.e., $\bar{U}^t \bar{U} = \bar{U} \bar{U}^t = I$).

b. Omit the $j$th variable (column) from $X$ and calculate the SVD of the remaining $n \times (p - 1)$ data matrix, denoted by $X_{-j} = \bar{U} \bar{D} \bar{V}$ with $\bar{U} = \{ \bar{u}_{-j} \}$, $\bar{V} = \{ \bar{v}_{-j} \}$ and $\bar{D} = diag(\bar{d}_1, ..., \bar{d}_{p-1})$.

c. Impute for $(i, j)$th missing case with

$$
\hat{x}_{ij} = \sum_{t=1}^{p-1} \bar{u}_{-j} \bar{d}_t^{1/2} \bar{v}_{-j} \bar{d}_t^{1/2}
$$

In the case where there is more than one missing value, an iterative scheme can be conducted as follows: start with any initial imputed values such as the mean, and update each initial imputed value in turn using $c$. The process is then iterated until a specified stopping rule is satisfied.

2.2. Variance estimation in complex survey designs

Horvitz-Thompson (1952) introduced an unbiased estimator for $T_X = \sum_{i=1}^{N} X_i$ for any design, with or without
replacement for \( i = 1, \ldots, N \). Let \( \pi_i \) be the probability that unit \( i \) is included in the sample under a given sampling scheme. The Horvitz-Thompson estimator is:

\[
\hat{T}_x = \sum_{i=1}^{n} \frac{X_i}{\pi_i}.
\]

Where \( n \) is the distinct number of units in the sample. The Horvitz-Thompson estimator does not depend on the number of times a unit may be selected. Each distinct unit of the sample is utilized only once. The following two formula are derived:

\[
E(\hat{T}_x) = T_x
\]

and

\[
V(\hat{T}_x) = \sum_{i=1}^{N} \left( \frac{1 - \pi_i}{\pi_i} \right) X_i^2 + \sum_{i=1}^{N} \sum_{j \neq i} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) X_i X_j
\]

where \( \pi_{ij} > 0 \) denotes the probability that both unit \( i \) and unit \( j \) are included.

The estimated variance of the Horvitz-Thompson estimator is given by:

\[
\hat{V}(\hat{T}_x) = \sum_{i=1}^{n} \left( \frac{1 - \pi_i}{\pi_i^2} \right) X_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{1}{\pi_{ij}} X_i X_j
\]

where \( \pi_{ij} > 0 \) denotes the probability that both unit \( i \) and unit \( j \) are included.

The unbiased estimator of the population mean of a variable \( X \) is given by

\[
\bar{X}_{HT} = \frac{\hat{T}_x}{N}.
\]

The variance of the Horvitz-Thompson estimator of \( \bar{X}_{HT} \), is given by

\[
V(\bar{X}_{HT}) = N^{-2} \times V(\hat{T}_x).
\]

Cluster sampling can be used in a single-stage design where the whole population of a cluster is recruited. From Lumley (2010), it is common for large surveys to report only cluster and strata information for the first stage of sampling, or even to report pseudo PSU and strata information. In the single-stage approximation the PSUs are treated as strata and the second-stage sampling units are treated as PSUs. The replicate-weight approach to estimating standard errors computes the standard deviation of the estimated summary across many partially independent subsets of the one sample. Replicate weights for single-stage cluster sampling designs are produced by treating the clusters as units. The jackknife estimates set the weight to 0 for a single observation and \( \frac{n}{(n-1)} \) for the other observation. For the variance estimations, we use jackknife estimator designed for stratified designs deleting one cluster at a time and use Bootstrap estimates taking a sample of clusters with replacement from each stratum and the sampling weight multiplied by the number of times the cluster appearing in the sample, see Yee et al. (1999).

The jackknife method, which was originally designed to estimate the bias of an estimator by deleting one datum from the original data set and recalculating the estimator based on the rest of the data, has become a valuable tool for the variance estimation since the work of Tukey (1958). In an infinite population context, Tukey suggested that each replicate estimate might be regarded as an independent and identically distributed random variable, which in turn suggests a very simple variance estimator. In the finite population sampling context, each jackknife replicate deletes one unit and modifies the weights of others. The general set up of the jackknife variance method is using the size of \( n \) samples, \( Y_1, Y_2, \ldots, Y_n \) for calculating \( \hat{\theta} = f(Y_1, Y_2, \ldots, Y_n) \), an estimator of parameter \( \theta \). Then similarly after deleting the \( k \)-th element, the mean estimator deleted one element can be obtained as follows

\[
\hat{\mu}_n(-k) = \frac{Y_1 + Y_2 + \ldots + Y_{k-1} + Y_{k+1} + \ldots + Y_n}{n-1}.
\]
3. Data set and simulation study

3.1. NHANES III data set

The National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC) collects, analyzes, and disseminates data on the health status of U.S. residents. The National Health and Nutrition Examination Survey (NHANES) is a periodic survey conducted by NCHS. The third National Health and Nutrition Examination Survey (NHANES III), conducted from 1988 through 1994, was the seventh in a series of these surveys based on a complex, multi-stage sample plan.

Because each sample person in the sample design does not have the same probability of selection, several aspects of the NHANES design must be taken into account in data analysis, including the sample weights and the complex survey design. Appropriate sample weights are needed to estimate prevalence, means, medians, and other statistics. The sample weights incorporate the differential probabilities of selection and include adjustments for noncoverage and nonresponse. For the survey description in detail, we recommend the website https://www.cdc.gov/nchs/data/nhanes/nhanes3/nh3gui.pdf.

The first stage of the design consisted of selecting a sample of 81 primary sampling unit (PSU)’s that were mostly individual counties. The PSU’s were stratified and selected with probability proportional to size (PPS).

The 89 locations were randomly divided into two groups, one for each phase. The first group consisted of 44 and the other of 45 locations. One set of PSU’s was allocated to the first three-year survey period (1988–91) and the other set to the second three-year period (1991–94). Therefore, unbiased estimates of health and nutrition characteristics can be independently produced for both Phase 1 and Phase 2 as well as for both phases combined.

We downloaded the NHANES III data with 17,030 observations and 16 variables from Hosmer and Lemeshow (Hosmer & Lemeshow, 2000)’s Textbook. A subset of data from the National Health and Nutrition Examination Study (NHANES) III. Subjects age $\geq 20$ are included. Forty-nine pseudo strata were created with 2 pseudo-PSU’s in each stratum. In this dataset, there are about 17,030 individuals, which are described by 16 variables. But 9,103 were excluded because they were missing one or more points for 16 variables. So 7,927 individuals are the number of complete dataset. In this paper, we treat 7,927 individuals as a sample for the population for the comparison and the relative efficiency for six imputation methods performed. This sample classifies people described by a set of attributes as whether smoking influence people or not i.e whether the individual got high blood pressure or not caused by smoking. In this research, we select five quantitative variables; Body Weight (BMPWTLBS), Serum Cholesterol (TCP), Standing Height (BMPHTIN), Average Systolic Blood Pressure (PEPMNK1R) and Average Diastolic BP (PEPMNK5R).

3.2. Simulation setting

In our simulation study, we select five quantitative variables; BMPWTLBS, BMPHTIN, PEPMNK1R, PEPMNK5R, and TCP from the “nhanes3” data set that is available in survey package of the statistical software R. There are 7,927 individuals who do not have any missing on the five variables. We use these as the complete data during our simulation study. Figures 1 and 2 show that three variables BMPWTLBS, PEPMNK1R, and TCP are skewed to the right. Table 2 summarizes the sample variance, skewness and excessive kurtosis of the five variables. We generate missing observations from these three variables at 5% and 15% missing completely at random rates. With 5% and 15% missing rates, the proportions of the complete observations are approximately 85.74% and 61.41%, respectively. For each configuration of the missing rate, we generated 500 incomplete datasets to compare and evaluate the performances of the imputation methods. Figure 3 shows typical missing patterns of the simulated datasets at 5% and 15% missing rates, respectively. For computing the variance estimation in complex survey data, we used survey R package from Complex Surveys: A Guide to Analysis Using R (Lumley, 2010).

4. Results

We summarize the results of the simulation studies in this section. Dong and Peng (2013) mentioned the proportion of missing data is directly related to the quality of statistical inferences. Yet, there is no established cutoff from
Table 1
NHANES III data set

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Codes/values</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Respondent ID Number</td>
<td>SEQN</td>
<td></td>
</tr>
<tr>
<td>2 Pseudo-PSU 1, 2 SDPPSU6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Pseudo-stratum 01–49 SDPSTRA6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Statistical weight 225.93–139744.9 WTPFX6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Age Years HSAGEIR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Sex 0 = Female, 1 = Male</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Race 1 = White, 2 = Black, 3 = Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Body weight Pounds BMPWTLBS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Standing height Inches BMPHTIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Average systolic BP mm Hg PEPMNK1R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Average diastolic BP mm Hg PEPMNK5R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Has respondent smoked &gt; 100 cigarettes in life 1 = Yes, 2 = No HAR1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 Does respondent smoke cigarettes now? 1 = Yes, 2 = No HAR3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 Smoking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 = if HAR1 = 1 &amp; HAR3 = 2 SMOKE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 = if HAR1 = 1 &amp; HAR3 = 1 SMOKE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 Serum cholesterol mg/100 ml TCP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 High blood pressure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 if PEPMNK1R ≤ 140 HBP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 if PEPMNK1R &gt; 140 HBP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. Histograms of the five variables: BMPWTLBS, BMPHTIN, PEPMNK1R, PEPMNK5R, and TCP.
Table 2
Sample variance, skewness and excessive kurtosis of the five variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variance</th>
<th>Skewness</th>
<th>Excessive Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMPWTLBS</td>
<td>1477.49</td>
<td>0.9721</td>
<td>2.3613</td>
</tr>
<tr>
<td>BMPHTIN</td>
<td>13.59</td>
<td>−0.0260</td>
<td>−0.2156</td>
</tr>
<tr>
<td>PEPMNK1R</td>
<td>375.12</td>
<td>0.9671</td>
<td>1.3511</td>
</tr>
<tr>
<td>PEPMNK5R</td>
<td>114.50</td>
<td>0.2865</td>
<td>0.9507</td>
</tr>
<tr>
<td>TCP</td>
<td>1922.61</td>
<td>0.8163</td>
<td>3.2284</td>
</tr>
</tbody>
</table>

Table 3
Average computing time of imputation methods in seconds

<table>
<thead>
<tr>
<th>Method</th>
<th>kNN</th>
<th>PCA</th>
<th>SVD</th>
<th>PMM</th>
<th>RF</th>
<th>Copula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>0.0246</td>
<td>0.423</td>
<td>0.736</td>
<td>4.956</td>
<td>109.588</td>
<td>576.033</td>
</tr>
</tbody>
</table>

Fig. 2. Q-Q plots of the five variables: BMPWTLBS, BMPHTIN, PEPMNK1R, PEPMNK5R, and TCP.

the literature regarding an acceptable percentage of missing data in a data set for valid statistical inferences. For example, Schafer (1999) asserted that a missing rate of 5% or less is inconsequential. Bennett (2001) maintained
Table 4
Simulation results: Standard errors of the sample mean at 5% missing rate

<table>
<thead>
<tr>
<th>Method</th>
<th>SRS</th>
<th>Stratified</th>
<th>Bootstrap</th>
<th>Jackknife</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMPWTLBS</td>
<td>0.7129</td>
<td>0.7128</td>
<td>0.7077</td>
<td>0.7130</td>
<td>1.3687</td>
</tr>
<tr>
<td>RF</td>
<td>0.7638</td>
<td>0.7636</td>
<td>0.7741</td>
<td>0.7639</td>
<td>1.1882</td>
</tr>
<tr>
<td>kNN</td>
<td>0.7150</td>
<td>0.7148</td>
<td>0.7133</td>
<td>0.7150</td>
<td>1.1928</td>
</tr>
<tr>
<td>PCA</td>
<td>0.7008</td>
<td>0.7005</td>
<td>0.7057</td>
<td>0.7007</td>
<td>1.2175</td>
</tr>
<tr>
<td>SVD</td>
<td>0.7151</td>
<td>0.7150</td>
<td>0.7056</td>
<td>0.7153</td>
<td>1.3358</td>
</tr>
<tr>
<td>PMM</td>
<td>0.7145</td>
<td>0.7144</td>
<td>0.7010</td>
<td>0.7146</td>
<td>1.3292</td>
</tr>
<tr>
<td>PEPMNK1R</td>
<td>0.2707</td>
<td>0.2703</td>
<td>0.2616</td>
<td>0.2704</td>
<td>0.0400</td>
</tr>
<tr>
<td>Copula</td>
<td>0.3210</td>
<td>0.3207</td>
<td>0.3105</td>
<td>0.3208</td>
<td>0.2762</td>
</tr>
<tr>
<td>RF</td>
<td>0.2649</td>
<td>0.2644</td>
<td>0.2608</td>
<td>0.2645</td>
<td>0.0828</td>
</tr>
<tr>
<td>kNN</td>
<td>0.2659</td>
<td>0.2654</td>
<td>0.2669</td>
<td>0.2655</td>
<td>0.1384</td>
</tr>
<tr>
<td>PCA</td>
<td>0.2632</td>
<td>0.2627</td>
<td>0.2625</td>
<td>0.2628</td>
<td>0.1325</td>
</tr>
<tr>
<td>SVD</td>
<td>0.2740</td>
<td>0.2736</td>
<td>0.2749</td>
<td>0.2737</td>
<td>0.0677</td>
</tr>
<tr>
<td>PMM</td>
<td>0.2742</td>
<td>0.2738</td>
<td>0.2716</td>
<td>0.2738</td>
<td>0.1085</td>
</tr>
<tr>
<td>TCP</td>
<td>0.6984</td>
<td>0.6981</td>
<td>0.7053</td>
<td>0.6983</td>
<td>1.7878</td>
</tr>
<tr>
<td>Copula</td>
<td>0.8763</td>
<td>0.8760</td>
<td>0.8835</td>
<td>0.8762</td>
<td>0.4353</td>
</tr>
<tr>
<td>RF</td>
<td>0.6867</td>
<td>0.6864</td>
<td>0.6790</td>
<td>0.6866</td>
<td>1.6583</td>
</tr>
<tr>
<td>kNN</td>
<td>0.7427</td>
<td>0.7426</td>
<td>0.7418</td>
<td>0.7429</td>
<td>1.8113</td>
</tr>
<tr>
<td>PCA</td>
<td>0.6817</td>
<td>0.6814</td>
<td>0.6760</td>
<td>0.6816</td>
<td>1.5734</td>
</tr>
<tr>
<td>SVD</td>
<td>0.7031</td>
<td>0.7026</td>
<td>0.7121</td>
<td>0.7028</td>
<td>1.4507</td>
</tr>
<tr>
<td>PMM</td>
<td>0.6995</td>
<td>0.6993</td>
<td>0.6847</td>
<td>0.6995</td>
<td>1.7857</td>
</tr>
</tbody>
</table>

Fig. 3. Missing patterns at 5% (Left) and 15% (Right) missing rates.

that statistical analysis is likely to be biased when more than 10% of data are missing. Furthermore, the amount of missing data is not the sole criterion by which a researcher assesses the missing data problem. Tabachnick and Fidell (2001) posited that the missing data mechanisms and the missing data patterns have greater impact on research results than does the proportion of missing data.

The illustration of the multivariate imputation models using the NHANES III dataset, which are described by 16 variables classifies people described by a set of attributes as whether smoking influence people or not i.e whether the individual got high blood pressure or not caused by smoking. In this example the covariate variables are showing the health condition of the individuals. Three skewed distributed independent variables were taken to this analysis and these variables describe the status of the individuals in Table 1.

We conducted simulation studies using National Health and Nutrition Survey data considering 5% and 15% missing rates. Based on our 500 times resampling simulation study of standard errors of the sample mean in complex
survey designs, we can conclude that the principal component analysis (PCA) imputation method appears to outperform other imputation methods overall in Tables 2 and 3. But the copula imputation method shows relatively good performance in clustering sampling design and 5% missing rate where the data are skewed to right because the standard error of each missing data design by using copula imputation method is smaller than other imputation methods. For TCP variable when 5% missing rate, the distribution of TCP variable is heavily skewed to right. The SE value of the copula imputation is much smaller than the SEs of other imputation methods. In addition, the RF imputation method shows relatively good performance in bootstrap sampling design and 5% missing rate where the data are skewed to right.

The execution time of computing the SEs was related to the size of the dataset especially to the percentage of imputation rate. Especially, under the jackknife method under stratified sampling, it reaches around 300 mins on the largest dataset at the 15% rate of missing values.

The computation time of six imputation methods are quite different. For example, the kNN imputation method is almost immediate whereas the copula imputation method requires 8–10 minutes for completion of imputation. Table 3 summarizes the computing time of each imputation method.

Missing data are a part of almost all research, and there are several alternative ways to overcome the drawbacks they produced. It was previously observed that neutral and well-designed comparison studies in computational sciences are necessary to ensure that previously proposed methods work as expected in various situations and to establish standards and guidelines (Boulesteix et al. 2013).

However, only a few studies report an evaluation of existing imputation methods (Brock et al., 2008; Celton et al., 2010; Luengo et al., 2012). In the present study, we performed a neutral comparison of six imputation methods based on four real datasets of various sizes, under an MAR assumption. Validation of imputation results is an important step and we consequently considered two evaluation criteria: standard error (SE) and execution time. While much attention has been paid to the imputation accuracy measured by Root Mean Square Error (RMSE), some studies have examined the effect of imputation on high-level analyses such as unsupervised and supervised classification (Boulesteix et al., 2013; Wang et al., 2006), or the time of execution (Saunders et al., 2006). We found that the sample means by using all six multivariate imputation methods are slightly biased compared with the unbiased sample mean by the complete sample data. To see the magnitude of the gain in efficiency of using the multivariate imputation methods, we compute the percent relative efficiency (RE(%)) of the mean square error (MSE) of the sample mean by using one of six multivariate imputation methods introduced with respect to the MSE of sample
mean by the complete sample data as follows

$$RE(\%) = \frac{MSE(\bar{X}_C)}{MSE(\bar{X}_I)} \times 100\%,$$

where $MSE(\bar{X}_I)$ denotes the MSE of the sample mean by one of six multivariate Imputation methods and $MSE(\bar{X}_C)$ denotes the MSE of the sample mean obtained by complete sample data, and we know $MSE(\bar{X}) = \text{Var}(\bar{X}) + \text{Bias}(\bar{X})^2$. If the value of $RE(\%)$ is greater than 100, then the proposed estimator by using one of six multivariate imputation methods is more efficient than the estimator by the complete sample data.

As shown in Table 6 at 5% missing rate, the values of $RE(\%)$ for BMPWTLBS are more than 100 for the RF and PCA imputation methods and the most of values of $RE(\%)$ for PEPMNK1R variable are more than 100 for the PCA imputation method. We can say that for 5% missing rate, the estimators by the RF and PCA imputation methods are more efficient than the estimator by the complete sample data when the data has high skewness. Similarly, Table 7 at 15% missing rate shows that the values of $RE(\%)$ for BMPWTLBS variable are more than 100 for the
PCA imputation method, the values of RE(%) for TCP variables are more than 100 for the RF imputation method. We can say that for 15% missing rate and the estimators by the RF and PCA imputation methods are mostly more efficient than the estimator by the complete sample data. Especially, the multivariate imputation methods appears to be efficient in the cluster sampling design when the data has high skewness or excessive kurtosis at the 15% MAR rate.

5. Discussion and conclusions

We compared several multivariate imputation methods for missing data. This work revealed novel insights into the performance of variance estimation of the sample mean by a variety of modern multivariate imputation methods including copula imputation, random forest imputation, and k-nearest neighbors imputation methods in complex designs such as stratified sampling, cluster sampling and replication approach to variance estimation by using jackknife and bootstrap methods in stratified sampling. We conducted simulation studies using National Health and Nutrition Survey data considering 5% and 15% missing rates. Based on our 500 times resampling simulation study of MSE of the sample mean in complex survey designs, the PCA and RF imputation methods appear to outperform other imputation methods when the data has high skewness or excessive kurtosis at the 15% missing rate. In the clustering sampling design, most of the multivariate imputation methods performed well even if the estimators by the multivariate imputation methods have biased estimators. Our findings in this paper suggest that more research is necessary to understand the effects of the multivariate imputation methods in complex survey designs.

References


