A note on approximate Bayesian bootstrap imputation

BY J. K. KIM

Department of Statistics, Hankuk University of Foreign Studies, Yongin, 449-791, Korea
kimj@hufs.ac.kr

SUMMARY

The approximate Bayesian bootstrap is suggested by Rubin & Schenker (1986) as a way of generating multiple imputations when the original sample can be regarded as independently and identically distributed and the response mechanism is ignorable. We investigate the finite sample properties of the variance estimator when the approximate Bayesian bootstrap method is used and show that the bias is not negligible for moderate sample sizes. A modification of the method is proposed for reducing the bias of the variance estimator. The proposed method is asymptotically equivalent to the approximate Bayesian bootstrap method but has better finite sample properties.

Some key words: Cell mean model; Multiple imputation; Nonresponse; Variance estimation.

1. Introduction

Imputation is a common technique for handling incomplete observations by filling in the missing values. Multiple imputation, proposed by Rubin (1978), allows the data analyst to use standard techniques of analysis designed for complete data, while at the same time providing a method of estimating the uncertainty associated with the missing data. However, several authors have outlined limitations to the multiple imputation approach. Fay (1992, 1993) explained that the multiple imputation variance is a biased estimator for domains that are not part of the imputer’s model. Meng (1994) also discussed the implications of a conflict between the imputer’s and the analyst’s models. Kott (1995) investigated the validity of the multiple-imputation variance estimator for samples with unequal weights. Binder & Sun (1996) investigated the validity of multiple imputation for general complex sampling schemes. Wang & Robins (1998) and Robins & Wang (2000) discussed the inconsistency of the multiple-imputation variance estimator and proposed new variance estimators, which require additional computation.

When the original sample can be regarded as independently and identically distributed, the approximate Bayesian bootstrap method, proposed by Rubin & Schenker (1986), seems to be well accepted as a method of multiple imputation. The method is an approximation to the Bayesian bootstrap, motivated by a Bayesian framework using Dirichlet process priors as uninformative priors. Binder (1982) and Lo (1986) showed the frequency validity of the Bayesian bootstrap for large samples.

Multiple imputation creates \( M > 1 \) repeated imputations for a dataset of size \( n \). Thus, when discussing asymptotic properties of the multiple imputation estimator, there are two different components to be considered, namely \( M \) and \( n \). Rubin & Schenker (1986) showed the consistency of the multiple-imputation variance estimator under the asymptotic set-up in which \( n \) and \( M \) go to infinity. For fixed \( M \) with a large value of \( n \), an interval estimator can be still constructed using a \( t \) reference distribution with associated degrees of freedom (Rubin, 1987, eqn (3.1.6)). However, the variance estimator itself is not consistent for fixed \( M \) (Wang & Robins, 1998). Frequency properties of the approximate Bayesian bootstrap imputation for fixed \( n \) were not discussed in Rubin & Schenker (1986). Recently, Barnard & Rubin (1999) noticed inappropriate behaviour of the
multiple-imputation interval estimator under modest sample sizes, and proposed an adjusted degrees of freedom for interval estimation, but their method is not applicable to variance estimation problems.

In this paper, under the frequentist paradigm, we show in § 2 that the approximate Bayesian bootstrap multiple-imputation variance estimator is biased for moderate sample sizes and the bias is always negative. In § 3 a simple modification of the method is proposed that reduces the bias of the variance estimator and preserves the asymptotic properties of the imputation. Limited simulations are presented in § 4.

2. Finite sample properties

Assume that the original sample is placed in $G$ cells. Let $n_g$ be the number of sample elements in imputation cell $g$ and let $r_g$ be the number of respondents in imputation cell $g$. Within cell $g$ ($g = 1, \ldots, G$), the elements in the sample are independently and identically distributed with mean $\mu_g$ and variance $\sigma_g^2$. We write this as

$$y_i \sim (\mu_g, \sigma_g^2),$$

independently for all $i \in A_{Rg}$, and

$$y_i \sim (\mu_g, \sigma_g^2),$$

independently for all $i \in A_{Mg}$, where $A_{Rg}$ denotes the set of indices for the responding units in the $g$th imputation cell, and $A_{Mg}$ denotes the set of indices for the nonresponding units in the $g$th imputation cell. We refer to the model defined by (1) and (2) as the cell mean model.

The approximate Bayesian bootstrap imputation can be described as follows.

**Step 1.** For each repetition of the imputation ($k = 1, \ldots, M$), select $Y^*_{g(k)} = \{y^*_{1(k)}, \ldots, y^*_{r(k)}\}$ with replacement with equal probabilities of selection, from the respondents $Y_{Rg} = \{y_i; i \in A_{Rg}\}$. The set of selected $Y^*_{g(k)}$ is called the donor set of cell $g$ at the $k$th repeated imputation.

**Step 2.** For each missing unit $j \in A_{Mg}$, draw $y^*_{j(k)}$ from the donor set $Y^*_{g(k)}$ of Step 1, again with replacement and with equal probabilities of selection, and use the selected value as the $k$th imputed value for unit $j \in A_{Mg}$.

**Step 3.** Repeat Steps 1 and 2 independently $M$ times.

To calculate the expectation of the approximate Bayesian bootstrap variance estimator, we need the following lemma. For the proof, see the Appendix.

**Lemma 1.** Assume that the first $r_g$ elements in cell $g$ respond and the remaining $m_g = n_g - r_g$ are missing. Let $y_i$ be the observed value of the $i$th unit, $i = 1, \ldots, r_g$, and let $y^*_{j(k)}$ be the imputed value associated with the $i$th unit with the $k$th repetition of the approximate Bayesian bootstrap method for $j = r_g + 1, \ldots, n_g$ and $k = 1, \ldots, M$. Then, under the cell mean model, for $i$ and $j$ in cell $g$,

$$\text{cov}(y_i, y^*_{j(k)}) = r_g^{-1}\sigma_g^2,$$

and

$$\text{cov}(y^*_{j(k)}, y^*_{j(k)}) = \begin{cases} \sigma_g^2 & \text{if } j = j', k = k', \\ (2r_g^{-1} - r_g^{-2})\sigma_g^2 & \text{if } j \neq j', k = k', \\ r_g^{-1}\sigma_g^2 & \text{if } k \neq k'. \end{cases}$$

Let $\hat{\theta}_n$ be the complete-data estimator of a parameter $\theta$ and let $\hat{V}_n$ be the complete-data estimator of the variance of $\hat{\theta}_n$. Using the $k$th multiple imputation data, we calculate $\hat{\theta}_{r(k),n}$ and $\hat{V}_{r(k),n}$, where $\hat{\theta}_{r(k),n}$ and $\hat{V}_{r(k),n}$ are $\hat{\theta}_n$ and $\hat{V}_n$ computed from the $k$th imputation dataset, respectively. Rubin (1978) proposed

$$\hat{\theta}_{M,n} = M^{-1} \sum_{k=1}^M \hat{\theta}_{r(k),n}.$$
as a point estimator of \( \theta \). The proposed estimator for the variance of \( \hat{\theta}_{M,n} \) can be written as
\[
\hat{V}_{M,n} = W_{M,n} + (1 + M^{-1})B_{M,n},
\]
where
\[
W_{M,n} = M^{-1} \sum_{k=1}^{M} \hat{V}_{i(k),n}, \quad B_{M,n} = \frac{1}{M-1} \sum_{k=1}^{M} (\hat{\theta}_{i(k),n} - \hat{\theta}_{M,n})^2.
\]

We consider the case when \( \hat{\theta}_{i,n} = n^{-1} \sum_{i=1}^{n} y_i \) is the complete-data point estimator and \( \hat{V}_{n} = n^{-1}(n-1)^{-1} \sum_{i=1}^{n} (y_i - \hat{\theta}_{n})^2 \) is the variance estimator for the complete sample. Using (3) and (4), we have, up to the order \( n^{-2} \) terms,
\[
E(W_{M,n}) = \text{var} \left( n^{-1} \sum_{g=1}^{G} n_{g} \mu_{g} \right) + E \left( n^{-2} \sum_{g=1}^{G} \frac{m_{g} n_{g}}{r_{g}} \left( 1 - \frac{m_{g}}{r_{g}} \right) \sigma_{g}^2 \right), \quad (5)
\]
\[
E(B_{M,n}) = E \left( n^{-2} \sum_{g=1}^{G} m_{g} \left( \frac{n_{g}}{r_{g}} - \frac{1}{r_{g}} - \frac{n_{g}}{r_{g}^3} \right) \sigma_{g}^2 \right). \quad (6)
\]

Therefore, the overall expectation of the variance estimator is, up to the order \( n^{-2} \) terms,
\[
E(\hat{V}_{M,n}) = n^{-2} \left[ \text{var} \left( \sum_{g=1}^{G} n_{g} \mu_{g} \right) + \sum_{g=1}^{G} E \left( \frac{n_{g}^2}{r_{g}} - \frac{m_{g} n_{g}}{r_{g}} \left( \frac{2}{n} + \frac{1}{n_{g}} \right) \right) \sigma_{g}^2 \right].
\]

The true variance of the multiply imputed estimator is, up to the order \( n^{-2} \) terms,
\[
\text{var}(\hat{\theta}_{M,n}) = \text{var} \left( n^{-1} \sum_{g=1}^{G} n_{g} \mu_{g} \right) + E \left( n^{-2} \sum_{g=1}^{G} \frac{n_{g}^2}{r_{g}} \sigma_{g}^2 \right)
+ M^{-1} E \left( n^{-2} \sum_{g=1}^{G} m_{g} \left( \frac{n_{g}}{r_{g}} - \frac{1}{r_{g}} - \frac{n_{g}}{r_{g}^3} \right) \sigma_{g}^2 \right). \quad (7)
\]

Derivations for (5), (6) and (7) are given in the Appendix. Therefore, the bias of the multiple-imputation variance estimator is, up to the order \( n^{-2} \) terms,
\[
E(\hat{V}_{M,n}) - \text{var}(\hat{\theta}_{M,n}) = -n^{-2} \left\{ \sum_{g=1}^{G} \frac{m_{g} n_{g}}{r_{g}} \left( \frac{2}{n} + \frac{1}{n_{g}} \right) \sigma_{g}^2 \right\}. \quad (8)
\]

The bias term is always negative.

Note also that the bias is independent of \( M \). This is demonstrated in Table 1, in that the absolute value of the relative bias of the multiple-imputation variance estimator is larger for large \( M \), because

<table>
<thead>
<tr>
<th>( n )</th>
<th>Resp. rate</th>
<th>( \text{ABB} \ (M = 3) )</th>
<th>( \text{ABB} \ (M = 10) )</th>
<th>( \text{ABB} \ (M = \infty) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.8</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-0.8</td>
</tr>
<tr>
<td>0.6</td>
<td>-1.7</td>
<td>-1.8</td>
<td>-1.8</td>
<td>-1.9</td>
</tr>
<tr>
<td>0.4</td>
<td>-2.8</td>
<td>-3.1</td>
<td>-3.3</td>
<td>-3.3</td>
</tr>
<tr>
<td>20</td>
<td>0.8</td>
<td>-4.0</td>
<td>-4.2</td>
<td>-4.2</td>
</tr>
<tr>
<td>0.6</td>
<td>-8.4</td>
<td>-9.0</td>
<td>-9.3</td>
<td>-9.3</td>
</tr>
<tr>
<td>0.4</td>
<td>-14.2</td>
<td>-15.7</td>
<td>-16.5</td>
<td>-16.5</td>
</tr>
</tbody>
</table>

Resp. rate, response rate; \( \text{ABB} \), approximate Bayesian bootstrap.
the denominator of the relative bias term is smaller for large $M$ while the numerator term remains
the same. The absolute value of the bias is larger for lower response rate by the bias formula in (8). Table 1 shows that the relative bias is not negligible for a sample of size $n = 20$.

3. Modified approximate Bayesian bootstrap method

The approximate Bayesian bootstrap method can be briefly described as using the bootstrap method
twice, first for selecting the donor set $Y_{g,i}^{*}$ from the respondents $Y_{g}$ and secondly for selected imputed
values $y_{j}^{**}$ ($j = r_g + 1, \ldots, n_g$) from the selected donor set $Y_{g,i}^{*}$. The size of the donor set in a cell is
equal to the number of respondents in the cell. In the modified approximate Bayesian bootstrap
method, we determine the size of donor set in cell $g$, denoted by $d_g$, to minimise the bias.

The following lemma provides the covariance structure of the imputed values based on the
modified method. For the proof, see the Appendix.

**Lemma 2.** Under the set-up of Lemma 1, assume a modified approximate Bayesian bootstrap method
in which the donor set contains $d_g$ elements. Then, under the cell mean model, we have, for $i$ and $j$ in
cell $g$,

$$
cov(y_i, y_{j(k)}^{**}) = r_g^{-1} \sigma_g^2, \quad \text{if } j = j', k = k',
$$

$$
cov(y_{j(k)}^{**}, y_{j'(k')}^{**}) = \begin{cases} 
\sigma_g^2 & \text{if } j \neq j', k = k', \\
gr_g^{-1} d_g^{-1} (r_g^{-1} + d_g^{-1}) \sigma_g^2 & \text{if } j = j', k \neq k', \\
r_g^{-1} \sigma_g^2 & \text{if } k = k', 
\end{cases} \quad \text{if } r_g^n
$$

where $r_g$ is the number of respondents and $d_g$ is the size of the donor set in cell $g$.

Using (9) and (10), we have

$$
E(W_{M,n}) = \text{var} \left( n^{-1} \sum_{g=1}^{G} n_g \bar{y}_g \right) + E \left( n^{-2} \sum_{g=1}^{G} n_g \left[ m_g - \frac{m_g}{n-1} \{2 + (m_g - 1)(r_g^{-1} + d_g^{-1} - r_g^{-1} d_g^{-1})\} \right] \sigma_g^2 \right), \quad (11)
$$

$$
E(B_{M,n}) = E \left( n^{-2} \sum_{g=1}^{G} m_g \left( 1 + \frac{m_g}{d_g} - \frac{1}{r_g} - \frac{n_g - 1}{r_g d_g} \right) \sigma_g^2 \right). \quad (12)
$$

Also, the true variance of the multiply imputed estimator is, up to order $n^{-2}$ terms,

$$
\text{var}(\hat{\theta}_{M,n}) = \text{var} \left( n^{-1} \sum_{g=1}^{G} n_g \bar{y}_g \right) + E \left( n^{-2} \sum_{g=1}^{G} \frac{n_g^2}{r_g} \sigma_g^2 \right) + M^{-1} \bar{E} \left( n^{-2} \sum_{g=1}^{G} m_g \left( 1 + \frac{m_g}{d_g} - \frac{1}{r_g} - \frac{n_g - 1}{r_g d_g} \right) \sigma_g^2 \right). \quad (13)
$$

Note that the right-hand sides of (12) and (13) are decreasing functions of the donor set size $d_g$.
Hence, decreasing $d_g$ increases the expectation of the multiple-imputation variance estimator and
the variance of the multiple-imputation point estimator. The increase in the variance can be reduced
by increasing the number of repeated imputations, $M$.

The bias of the multiple-imputation variance estimator is

$$
E(\hat{\theta}_{M,n}) - \text{var}(\hat{\theta}_{M,n}) = -n^{-2} \left[ \sum_{g=1}^{G} \frac{m_g}{r_g d_g} \left( 2r_g d_g - n - 1 \right) + \frac{m_g - 1}{n - 1} \{r_g + d_g - 1\} \right] \sigma_g^2,
$$

$$
- r_g m_g + d_g (m_g + 1) + n_g - 1 \right) \sigma_g^2 \right]. \quad (14)
$$
The right-hand side of (14) is equal to zero when

\[ d_g = \frac{(r_g - 1)(m_g - 1)(n - 2)}{(n - 1)(m_g + 1) + n_g + r_g - 1}. \]  

(15)

Note that we always have \( d_g < r_g \) and \( d_g \to r_g \) as \( n \to \infty \), which justifies the choice \( d_g = r_g \) of Rubin & Schenker (1986) for large samples. For a moderate sample size, the donor set size \( d_g \) in (15) will lead to an unbiased multiple-imputation variance estimator, for any fixed \( M(M > 1) \). In practice, we choose the nearest integer value to the \( d_g \) of (15), greater than one, to implement the modified approximate Bayesian bootstrap imputation. The idea of modifying the bootstrap sample size also appears in Rao & Wu (1988) in the context of bootstrap variance estimation for sampling from a stratified random sample.

4. Simulation study

To test our theory, we performed a limited simulation. The simulation study can be described as a \( 2 \times 2 \times 3 \times 2 \times 2 \) factorial design with \( B = 10000 \) samples within each cell. The factors are as follows: two sampling distributions, \( N(5, 1) \) and \( \chi^2_5 \); two methods of multiple imputation, namely approximate Bayesian bootstrap and modified approximate Bayesian bootstrap; three response rates, \( p \), namely 0.8, 0.6 and 0.4; two sample sizes, \( n \), namely 20 and 100; and two numbers of repeated imputations, \( M \), namely 3 and 10. We assume a single imputation cell and consider the estimation of the population mean. The number of realised respondents is the same for each sample. In the modified method of § 3, we choose the size of the donor set to be the largest integer less than or equal to the value in (15).

The mean and variance of the point estimators and the relative bias of the estimators of variance are calculated. The point estimators of the population mean are unbiased under the two imputation schemes and are not listed here. Table 2 presents the variances of the point estimators and the relative biases of the two variance estimators under the two different imputation schemes. The relative bias of \( \hat{V} \) as an estimator of the variance of \( \hat{y}_I \) is calculated as

\[ \frac{\text{var}_B(\hat{y}_I)}{\text{var}_B(\hat{y}_I)} \left\{ E_B(\hat{V}) - \text{var}_B(\hat{y}_I) \right\}, \]

where the subscript \( B \) denotes the distribution generated by the Monte Carlo simulation.

The following remarks can be made on the basis of Table 2.

Remark 1. There are differences in the efficiency of the procedures. In small samples (\( n = 20 \)), the approximate Bayesian bootstrap procedure is slightly more efficient than the modified procedure. This is because, in the latter, the size of the donor set is smaller than the number of the respondents, according to formula (15). The fact that a smaller donor set is used in the modified method reduces the efficiency of the resulting point estimator relative to the approximate Bayesian bootstrap. The phenomenon is not obvious for large samples, such as \( n = 100 \), because the size of the donor set converges to the number of respondents when \( n \) is sufficiently large.

Remark 2. The relative bias of the variance estimator under the approximate Bayesian bootstrap imputation is consistent with the theoretical results; the relative bias is negative and its absolute value is larger for smaller samples and for smaller response rate.

Remark 3. The modified approximate Bayesian bootstrap imputation produces much smaller relative bias for the variance estimator. The main source of bias in the corresponding variance estimator is the fact that we have to choose an integer value for the donor size. For example, under \( M = 10 \), the relative bias is largest for \( n = 20 \) and \( p = 0.6 \), where the optimal value is 6·86. Thus, the fact that we choose the donor set size to be 6 instead of 6·86 makes the bias relatively large. In that case, one may take the donor set size 6 with probability 0·14 and take the donor set size 7 with probability 0·86, if a smaller bias is desired.
**Table 2. Simulation study. Variances of the point estimators and percentage relative biases of the variance estimators under the two different imputation schemes, based on 10,000 samples**

<table>
<thead>
<tr>
<th>M</th>
<th>Imputation method</th>
<th>Sample size, n</th>
<th>Response rate, p</th>
<th>Variance Normal $\chi^2$</th>
<th>Relative bias (%) Normal $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ABB</td>
<td>20</td>
<td>0.8</td>
<td>0.0654 0.652</td>
<td>-2.68 -2.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.0930 0.954</td>
<td>-7.70 -8.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.1471 1.442</td>
<td>-13.89 -11.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>0.8</td>
<td>0.0133 0.133</td>
<td>-0.26 -0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.0191 0.190</td>
<td>-2.55 -2.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.0298 0.298</td>
<td>-3.05 -2.96</td>
</tr>
<tr>
<td>3</td>
<td>MABB</td>
<td>20</td>
<td>0.8</td>
<td>0.0672 0.656</td>
<td>0.40 2.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.0970 1.001</td>
<td>2.70 -0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.1528 1.531</td>
<td>2.20 1.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>0.8</td>
<td>0.0135 0.134</td>
<td>-1.43 -0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.0191 0.189</td>
<td>-1.03 0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.0301 0.301</td>
<td>1.20 0.87</td>
</tr>
<tr>
<td>10</td>
<td>ABB</td>
<td>20</td>
<td>0.8</td>
<td>0.0633 0.639</td>
<td>-3.18 -4.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.0866 0.869</td>
<td>-9.73 -9.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.1293 1.312</td>
<td>-13.70 -14.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>0.8</td>
<td>0.0127 0.129</td>
<td>-0.25 -2.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.0174 0.174</td>
<td>-1.75 -2.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.0265 0.267</td>
<td>-2.82 -4.28</td>
</tr>
<tr>
<td>10</td>
<td>MABB</td>
<td>20</td>
<td>0.8</td>
<td>0.0638 0.645</td>
<td>0.69 -0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.0876 0.886</td>
<td>1.93 1.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.1321 1.336</td>
<td>1.35 0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>0.8</td>
<td>0.0127 0.129</td>
<td>0.35 -1.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.0173 0.172</td>
<td>0.72 0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
<td>0.0265 0.269</td>
<td>1.18 -0.64</td>
</tr>
</tbody>
</table>

ABB, approximate Bayesian bootstrap; MABB, modified approximate Bayesian bootstrap.

**Acknowledgement**

This paper was written while the author was working at Westat, Rockville, MD, U.S.A. The research was partially supported by a grant from the U.S. National Center for Education Statistics. The author thanks Wayne Fuller, Aref Dajani, Hyunshik Lee and Lou Lizzo for helpful comments. The comments from the referees, including the editor, greatly improved the presentation of the paper.

**Appendix**

**Proofs**

Proofs of Lemma 1 and Lemma 2. Since Lemma 1 is a special case of Lemma 2 with $d_g = r_g$, we prove only Lemma 2. Let $Y_{Rg}$ be the set of respondents and let $Y_{g(k)}^{*}$ be the donor set of size $d_g$ for the $k$th repetition of the modified approximate Bayesian bootstrap imputation in cell $g$. The choice $d_g = r_g$ will lead to the original approximate Bayesian bootstrap imputation of Rubin & Schenker (1986). Thus, in the $k$th repetition of the modified approximate Bayesian bootstrap imputation, we first select $Y_{g(k)}^{*}$ of size $d_g$ from $Y_{Rg}$ and then select $y_{j(k)}^{*}$, for $j = r_g + 1, \ldots, n_g$. For $i = 1, \ldots, r_g$ and $j = r_g + 1, \ldots, n_g$,
Define the augmented data at the cases of those for (11), (12) and (13) with

\[
\text{cov}(y_i, y_{j \mid Y_{Rg}}^*) = \text{cov}(y_i, E(y_{j \mid Y_{Rg}}^*)) = \text{cov}(y_i, \tilde{y}_{Rg}) = r_g^{-1} \sigma_g^2.
\]

Let \( \tilde{y}_{g(k)}^* \) and \( s_{g(k)}^2 \) be the sample mean and the sample variance of the set \( Y_{g(k)}^* \), respectively. Also, let \( \bar{y}_{Rg} \) and \( s_{Rg}^2 \) be the sample mean and the sample variance of the set \( Y_{Rg} \), respectively. Note that

\[
\text{var}(\tilde{y}_{g(k)}^*) = \text{var}(E(\tilde{y}_{g(k)}^* \mid Y_{Rg})) + E(\text{var}(\tilde{y}_{g(k)}^* \mid Y_{Rg}))
\]

\[
= \text{var}(\bar{y}_{Rg}) + E(d_g^{-1}(1 - r_g^{-1})s_{Rg}^2) = (r_g^{-1} + d_g^{-1} - r_g^{-1}d_g^{-1})\sigma_g^2.
\]

and, for \( k \neq k' \),

\[
\text{cov}(\tilde{y}_{g(k)}^*, \tilde{y}_{g(k')}^*) = \text{cov}(E(\tilde{y}_{g(k)}^* \mid Y_{Rg}), E(\tilde{y}_{g(k')}^* \mid Y_{Rg})) + E(\text{cov}(\tilde{y}_{g(k)}^*, \tilde{y}_{g(k')}^* \mid Y_{Rg}))
\]

\[
= \text{var}(\bar{y}_{Rg}) - r_g^{-1}\sigma_g^2.
\]

Also,

\[
E(s_{g(k)}^2) = E((1 - r_g^{-1})s_{Rg}^2) = (1 - r_g^{-1})\sigma_g^2.
\]

To show (10), first note that the \( y_{g(k)}^* \)s are distributed independently, conditional on \( Y_{g(k)}^* \). Therefore, for \( k \neq k' \),

\[
\text{cov}(y_{g(k)}^*, y_{j \mid Y_{g(k)}}^*) = \text{cov}(E(y_{g(k)}^* \mid Y_{g(k)}), E(y_{j \mid Y_{g(k)}}^*)) = \text{cov}(y_{g(k)}, y_{j \mid Y_{g(k)}}) = r_g^{-1}\sigma_g^2.
\]

For each \( k \), we have

\[
\text{var}(y_{g(k)}^*) = \text{var}(E(y_{g(k)}^* \mid Y_{g(k)})) + E(\text{var}(y_{g(k)}^* \mid Y_{g(k)}))
\]

\[
= \text{var}(\bar{y}_{g(k)}^*) + E((1 - d_g^{-1})s_{g(k)}^2) = \sigma_g^2.
\]

and, for \( j \neq j' \),

\[
\text{cov}(y_{g(k)}^*, y_{j \mid Y_{g(k)}}^*) = \text{cov}(E(y_{g(k)}^* \mid Y_{g(k)}), E(y_{j \mid Y_{g(k)}}^*)) = \text{cov}(y_{g(k)}^*, y_{j \mid Y_{g(k)}}) = r_g^{-1}\sigma_g^2.
\]

\[
\boxdot
\]

Proof of equations (5), (6), (7) and (11), (12), (13). The proofs for (5), (6) and (7) are special cases of those for (11), (12) and (13) with \( d_g = r_g \). It is therefore enough to establish (11), (12) and (13). Define the augmented data at the \( k \)th repeated imputation as

\[
\eta_{i(k)} = \begin{cases} 
    y_i & (i = 1, \ldots, r_g), \\
    y_{g(k)}^* & (i = r_g + 1, \ldots, n_g).
\end{cases}
\]

Then, we can write

\[
\hat{V}_n = n^{-1}(n-1)^{-1} \left( \sum_{i=1}^n y_i^2 - n\bar{y}_n^2 \right),
\]

\[
\hat{V}_{1(k),n} = n^{-1}(n-1)^{-1} \left( \sum_{i=1}^n \eta_{i(k)}^2 - n(\bar{\eta}_{1(k),n})^2 \right).
\]

Let \( N = \{n_1, \ldots, n_\alpha\} \) and \( R = \{r_1, \ldots, r_\alpha\} \). By Lemma 2, we have

\[
E(\hat{\theta}_{1(k),n} \mid N, R) = \sum_{g=1}^{\alpha} n_g \mu_g = E(\hat{\theta}_n \mid N, R),
\]

\[
E \left( \sum_{i=1}^n \eta_{i(k)}^2 \mid N, R \right) = \sum_{g=1}^{\alpha} n_g \mu_g^2 + \sum_{g=1}^{\alpha} n_g \sigma_g^2 = E \left( \sum_{i=1}^n y_i^2 \mid N, R \right).
\]

Hence,

\[
E(\hat{V}_{1(k),n}) = E(\hat{V}_n) + (n-1)^{-1}(\text{var}(\hat{\theta}_n) - \text{var}(\hat{\theta}_{1(k),n})) \cdot (A1)
\]
Under model (1) and (2),
\[ \text{var}(\hat{\theta}_g) = \text{var}\left( n^{-1} \sum_{g=1}^{G} n_g \mu_g \right) + E\left( n^{-2} \sum_{g=1}^{G} n_g \sigma_g^2 \right). \]

Using (9) and (10), we have
\[ \text{var}(\hat{\theta}_{I(k),n}) = \text{var}(\hat{\theta}_n) + E\left( n^{-2} \sum_{g=1}^{G} \left[ m_g^2 \left( 2 + (m_g - 1)(r_g^{-1} + d_g^{-1} - r_g^{-1}d_g^{-1}) \right) \right] \sigma_g^2 \right). \] (A2)

Since the \( M \) imputed estimators \( \hat{\theta}_{I(1),n}, \ldots, \hat{\theta}_{I(M),n} \) are identically distributed, we have \( E(W_{M,n}) = E(V_{I(n),n}) \). Therefore, inserting (A2) into (A1), we prove (11). To show (12), note that
\[ E(B_{M,n}) = \text{var}(\hat{\theta}_{I(k),n}) - \text{cov}(\hat{\theta}_{I(k),n}, \hat{\theta}_{I(k')},n). \]

For any \( k \neq k' \), using (9) and (10) again, we have
\[ \text{cov}(\hat{\theta}_{I(k),n}, \hat{\theta}_{I(k'),n}) = \text{var}\left( n^{-1} \sum_{g=1}^{G} n_g \mu_g \right) + E\left( n^{-2} \sum_{g=1}^{G} \left( r_g + 2m_g + \frac{m_g^2}{r_g} \right) \sigma_g^2 \right). \] (A3)

Therefore, we have
\[ E(B_{M,n}) = E\left( n^{-2} \sum_{g=1}^{G} m_g \left( 1 + \frac{m_g}{d_g} - \frac{1}{r_g} \frac{n_g - 1}{r_g d_g^2} \right) \sigma_g^2 \right). \]

Equally (13) is easily proved by inserting (A2) and (A3) into
\[ \text{var}(\hat{\theta}_{I(M),n}) = M^{-1} \text{var}(\hat{\theta}_{I(1),n}) + (1 - M^{-1}) \text{cov}(\hat{\theta}_{I(1),n}, \hat{\theta}_{I(k'),n}). \]

for any \( k \neq k' \).

**References**


[Received January 2001. Revised September 2001]