Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 4064–4068
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

# On the performance of bias–reduction techniques for variance estimation in approximate Bayesian bootstrap imputation

Hakan Demirtas[a,*], Lester M. Arguelles[a], Hwan Chung[b], Donald Hedeker[a]

[a] *Division of Epidemiology and Biostatistics (MC923), University of Illinois, 1603 West Taylor Street, Chicago, IL 60612, USA*
[b] *Department of Epidemiology, Michigan State University, B601 West Fee Hall East Lansing, MI 48824, USA*

Received 20 July 2006; received in revised form 25 December 2006; accepted 28 December 2006
Available online 30 January 2007

**Abstract**

Multiply imputed data sets can be created with the approximate Bayesian bootstrap (*ABB*) approach under the assumption of ignorable nonresponse. The theoretical development and inferential validity are predicated upon asymptotic properties; and biases are known to occur in small-to-moderate samples. There have been attempts to reduce the finite-sample bias for the multiple imputation variance estimator. In this note, we present an empirical study for evaluating the comparative performance of the two proposed bias–correction techniques and their impact on precision. The results suggest that to varying degrees, bias improvements are outweighed by efficiency losses for the variance estimator. We argue that the original *ABB* has better small-sample properties than the modified versions in terms of the integrated behavior of accuracy and precision, as measured by the root mean-square error.


*Keywords:* Multiple imputation; Variance estimation; Bootstrapping

## 1. Introduction

Missing data are the rule rather than the exception in statistical practice. Determining an appropriate analytical strategy in the absence of completeness is a consequential focus of scientific exploration on account of the extra intricacy that arises through missing data. Missing values generally sophisticate the statistical analysis in terms of biased parameter estimates, reduced statistical power and degraded confidence intervals, and thereby may lead to false inferences (Little and Rubin, 2002).

Improvements in computational statistics have produced flexible missing-data procedures with a sensible statistical basis. One of these procedures involves multiple imputation (*MI*) (Rubin, 1987), a simulation technique that replaces each missing datum with a set of $m > 1$ plausible values. The $m$ versions of complete data are then analyzed by standard complete-data methods and the results are combined into a single inferential summary using arithmetic rules to yield estimates, standard errors and $p$-values that formally consolidate missing data uncertainty into the modeling process. The key ideas and benefits of *MI* were reviewed by Rubin (1996) and Schafer (1997, 1999).

The fundamental step in *MI* is filling in the missing data by drawing from the conditional distribution of the missing data given the observed data which usually entails propounding a parametric model for the data and using it to derive

* Corresponding author. Tel.: +1 312 9969841; fax:+1 312 9960064.
*E-mail address:* demirtas@uic.edu (H. Demirtas).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.047

this conditional distribution. Rubin (1987) and Rubin and Schenker (1986) proposed a nonparametric *MI* technique, the approximate Bayesian bootstrap (*ABB*), which involves a two-stage sampling procedure as a way to generate proper imputations with minimal distributional assumptions for the intended sample (i.e., complete data). Consider the univariate sample $Y = y_1, \ldots, y_n$, where the first $r$ values are observed and the remaining $n-r$ values are missing. In the *ABB*, one creates a new pool of respondents by sampling $r$ values from the observed part of the data with replacement at the first stage. Subsequently, sampling $n - r$ values again with replacement from these "new" data leads to a set of imputations for the missing elements. The method, which is most appropriate for large samples, produces approximate draws from the conditional distribution of the missing data given the observed data under a multinomial model with categories corresponding to the distinct values seen in the observed data. Resampling at the second stage of the *ABB* approximates a draw of the multinomial probabilities from their observed-data posterior under a Dirichlet prior.

Unbiasedness of the mean and variance estimators is an asymptotic property under the *ABB*; however, the variance estimator is known to be biased in small-to-moderate samples. Quite recently, the two bias–reduction approaches have been proposed to address this problem (Kim, 2002 and Parzen et al., 2005). In this work, we empirically evaluate these suggested modifications relative to the classical *ABB* from an efficiency standpoint. The organization of the rest of this note is as follows. In the next section, we provide more details on the *ABB*, elaborate on small-sample bias and the suggested remedies. In Section 3, we present a study in which the data generation mechanisms leading to the simulated data are identical to the ones that appeared in the above-mentioned papers; this is done in an effort to assess the comparative performance on a broader spectrum that includes measures of precision. We conclude with some remarks and discussion.

## 2. Bias in variance estimation and suggested refinements

For what follows, the missingness mechanism is assumed to be ignorable in the sense that $(y_1, \ldots, y_r)$ is a random sample from $(y_1, \ldots, y_n)$ and each $y_i$, $i = 1, 2, \ldots, n$ is independently and identically distributed with mean $\mu$ and variance $\sigma^2$. For inferences regarding $\mu$, the *ABB* performs *MI* via the following steps:

1. Independently draw $r$ observations with replacement from the observed data, $(y_1, \ldots, y_r)$.
2. Draw the $n - r$ missing values with replacement from the $y_i$'s drawn in (1). The imputed values are denoted as $(\tilde{y}_{r+1}, \ldots, \tilde{y}_n)$
   - From Steps 1 and 2, the estimated mean is $\hat{\mu} = n^{-1}(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \tilde{y}_i)$, and the estimated within-imputation variance is $U = \widehat{V}ar(\hat{\mu}) = [n(n-1)]^{-1}[\sum_{i=1}^{r}(y_i - \hat{\mu})^2 + \sum_{i=r+1}^{n}(\tilde{y}_i - \hat{\mu})^2]$.
3. Repeat Steps 1 and 2 independently $m$ times. Analyses of the $m$ imputed data sets yield $\hat{\mu}^k$ and $U^k = \widehat{V}ar(\hat{\mu}^k)$, for $k = 1, \ldots, m$.

The *MI* estimator of $\mu$ is $\hat{\mu}_{\mathrm{MI}} = m^{-1}\sum_{k=1}^{m} \hat{\mu}^k$, and the variance estimator is $\hat{V} = \hat{W} + (\frac{m+1}{m})\hat{B}$, where $\hat{W} = m^{-1}\sum_{k=1}^{m} U^m$ is the average within-imputation variance, and $\hat{B} = (m-1)^{-1}\sum_{k=1}^{m}(\hat{\mu}^m - \hat{\mu}_{\mathrm{MI}})^2$ is the between-imputation variance. As $n \to \infty$ and $m \to \infty$, the *ABB* yields an unbiased estimator of $\mu$ and $Var(\hat{\mu})$ (Rubin, 1987). However, in the finite-sample case, $\hat{V}$ is always negatively biased by an amount given by $E(\hat{V}) - Var(\hat{\mu}_{\mathrm{MI}}) = -\frac{(n-r)}{n^2 r}(\frac{3}{n} + \frac{1}{r})\sigma^2$. In order to circumvent this, Kim (2002) proposed a modification of Step 1 of the *ABB* based on reducing the number of draws from $r$ to $d = \frac{(r-1)(n-r-1)(n-2)}{(n-1)(n-r+1)+n+r-1}$. Parzen et al. (2005) proposed to stick with the original sampling scheme and argued that the bias can be completely eliminated by multiplying $\hat{V}$ by a factor of $\frac{Var(\hat{\mu}_{\mathrm{MI}})}{E(\hat{V})}$, so that $E(\hat{V}) = Var(\hat{\mu}_{\mathrm{MI}})$. In both papers, the authors only considered improvements in accuracy with little mention of efficiency losses. However, reduction in bias is almost invariably associated with decreased precision due to the historical tradeoff between bias and variance. To more fully assess the impact of these approaches on the quality of variance estimators in comparison to the original *ABB*, we carried out a simulation study which is identical to the one that appeared in these papers, but with more comprehensive evaluation criteria that include efficiency measures.

## 3. A simulation study

The design of our simulation is the same as in Kim (2002) and Parzen et al. (2005). The complete data were generated following the standard normal distribution and $\chi^2$ distribution with five degrees of freedom. Missing values were imposed under a missing-completely-at-random mechanism. We considered a factorial design that includes two

Table 1
Results for small samples ($n = 20$)

| Distribution | Method | $r$ | $m$ | *TV* | *AE* | *RB1* | *RB2* | *RMSE* | *REL.EFF* |
|---|---|---|---|---|---|---|---|---|---|
| $N(0, 1)$ | *OABB* | 8 | 3 | 0.1456250 | 0.1268034 | −12.92 | −15.97 | 0.11933 | — |
| | *KIM* | | | | 0.1520466 | 4.41 | 4.36 | 0.14738 | 64.04 |
| | *PLF* | | | | 0.1477260 | 1.44 | 1.53 | 0.13729 | 73.68 |
| | *OABB* | 8 | 10 | 0.1311875 | 0.1123543 | −14.36 | −26.32 | 0.07399 | — |
| | *KIM* | | | | 0.1356790 | 3.42 | 5.06 | 0.08890 | 64.95 |
| | *PLF* | | | | 0.1333135 | 1.62 | 2.50 | 0.08493 | 71.03 |
| | *OABB* | 12 | 3 | 0.0929630 | 0.0850584 | −8.50 | −14.00 | 0.05700 | — |
| | *KIM* | | | | 0.0942244 | 1.38 | 1.87 | 0.06837 | 68.20 |
| | *PLF* | | | | 0.0928246 | −0.14 | −0.22 | 0.06160 | 83.97 |
| | *OABB* | 12 | 10 | 0.0862222 | 0.0784948 | −8.96 | −20.49 | 0.03848 | — |
| | *KIM* | | | | 0.0867991 | 0.67 | 1.33 | 0.04335 | 75.64 |
| | *PLF* | | | | 0.0861775 | 0.06 | 0.13 | 0.04144 | 82.77 |
| | *OABB* | 16 | 3 | 0.0661979 | 0.0636504 | −3.85 | −8.76 | 0.02918 | — |
| | *KIM* | | | | 0.0673362 | 1.72 | 3.39 | 0.03362 | 74.87 |
| | *PLF* | | | | 0.0663111 | 0.17 | 0.37 | 0.03029 | 92.14 |
| | *OABB* | 16 | 10 | 0.0636094 | 0.0607923 | −4.43 | −12.07 | 0.02350 | — |
| | *KIM* | | | | 0.0640127 | 0.64 | 1.61 | 0.02515 | 86.10 |
| | *PLF* | | | | 0.0634415 | −0.26 | −0.69 | 0.02435 | 91.82 |
| $\chi^2(5)$ | *OABB* | 8 | 3 | 1.4562500 | 1.2485900 | −14.26 | −15.09 | 1.39163 | — |
| | *KIM* | | | | 1.5519190 | 6.57 | 4.87 | 1.96694 | 49.06 |
| | *PLF* | | | | 1.4546070 | −0.11 | −0.10 | 1.60309 | 73.68 |
| | *OABB* | 8 | 10 | 1.3118750 | 1.1236890 | −14.34 | −19.24 | 0.99591 | — |
| | *KIM* | | | | 1.3395060 | 2.11 | 2.30 | 1.20087 | 66.36 |
| | *PLF* | | | | 1.3333090 | 1.63 | 1.85 | 1.16060 | 71.03 |
| | *OABB* | 12 | 3 | 0.9296296 | 0.8578892 | −7.17 | −9.89 | 0.72874 | — |
| | *KIM* | | | | 0.9593181 | 3.19 | 3.45 | 0.86205 | 70.85 |
| | *PLF* | | | | 0.9362182 | 0.71 | 0.83 | 0.79145 | 83.97 |
| | *OABB* | 12 | 10 | 0.8622222 | 0.7884803 | −8.55 | −13.91 | 0.53531 | — |
| | *KIM* | | | | 0.8718866 | 1.12 | 1.58 | 0.61183 | 75.12 |
| | *PLF* | | | | 0.8666582 | 0.51 | 0.76 | 0.58279 | 82.77 |
| | *OABB* | 16 | 3 | 0.6619792 | 0.6453358 | −2.54 | −3.96 | 0.42081 | — |
| | *KIM* | | | | 0.6775311 | 2.35 | 3.40 | 0.45814 | 84.33 |
| | *PLF* | | | | 0.6723130 | 1.56 | 2.36 | 0.43818 | 92.14 |
| | *OABB* | 16 | 10 | 0.6360938 | 0.6130286 | −3.63 | −6.84 | 0.33794 | — |
| | *KIM* | | | | 0.6425993 | 1.02 | 1.77 | 0.36682 | 84.50 |
| | *PLF* | | | | 0.6397435 | 0.57 | 1.04 | 0.35186 | 91.82 |

*TV*, *AE*, *RB*, *RMSE*, and *REL.EFF* stand for the true value, average estimate, relative bias, root mean-square error, and relative efficiency, respectively. $r$ and $m$ correspond to the number of observed values and the number of imputations, respectively. *OABB* stands for original approximate Bayesian bootstrap, and *KIM* and *PLF* represent the approach taken by Kim (2002) and Parzen et al. (2005), respectively. (The latter comes from the initials of the three authors.)

levels of $n$ (20 and 100), three levels of $r$ that correspond to 40%, 60%, and 80% response rates, two numbers of imputations ($m = 3, 10$). The experiment was repeated 10,000 times. We implemented the original *ABB* on these simulated data sets as well as the refinements proposed by Kim (2002) and Parzen et al. (2005).

Our evaluation criteria consist of the following quantities: the standardized bias (*SB*) is the relative magnitude of the raw bias with respect to the overall uncertainty in the system. In this context, it is equal to $100 \times \frac{E(\hat{V})-V}{SE(\hat{V})}$, where $V = Var(\hat{\mu}_{MI})$ and *SE* stands for standard error. The percentage bias (*PB*) is the ratio of the raw bias and the true value

Table 2
Results for moderate samples ($n = 100$)

| Distribution | Method | $r$ | $m$ | $TV$ | $AE$ | $RB1$ | $RB2$ | $RMSE$ | $REL.EFF$ |
|---|---|---|---|---|---|---|---|---|---|
| $N(0, 1)$ | *OABB* | 40 | 3 | 0.0298250 | 0.0291153 | −2.38 | −3.34 | 0.02122 | — |
| | *KIM* | | | | 0.0297950 | −0.10 | −0.14 | 0.02164 | 96.04 |
| | *PLF* | | | | 0.0299435 | 0.39 | 0.54 | 0.02181 | 94.54 |
| | *OABB* | 40 | 10 | 0.0264475 | 0.0255648 | −3.34 | −8.90 | 0.00985 | — |
| | *KIM* | | | | 0.0262253 | −0.73 | −1.93 | 0.01006 | 95.21 |
| | *PLF* | | | | 0.0263879 | −0.23 | −0.59 | 0.01013 | 93.86 |
| | *OABB* | 60 | 3 | 0.0188296 | 0.0184854 | −1.83 | −3.56 | 0.00968 | — |
| | *KIM* | | | | 0.0190213 | 1.02 | 1.86 | 0.01029 | 88.45 |
| | *PLF* | | | | 0.0187960 | −0.18 | −0.34 | 0.00984 | 96.72 |
| | *OABB* | 60 | 10 | 0.0173156 | 0.0168977 | −2.41 | −8.98 | 0.00467 | — |
| | *KIM* | | | | 0.0173560 | 0.23 | 0.82 | 0.00491 | 89.98 |
| | *PLF* | | | | 0.0172068 | −0.63 | −2.29 | 0.00474 | 96.44 |
| | *OABB* | 80 | 3 | 0.0133146 | 0.0133027 | −0.09 | −0.29 | 0.00414 | — |
| | *KIM* | | | | 0.0133285 | 0.10 | 0.33 | 0.00416 | 99.11 |
| | *PLF* | | | | 0.0133410 | 0.71 | 1.28 | 0.00417 | 98.41 |
| | *OABB* | 80 | 10 | 0.0127444 | 0.0126524 | −0.72 | −3.76 | 0.00245 | — |
| | *KIM* | | | | 0.0127649 | 0.16 | 0.83 | 0.00248 | 97.88 |
| | *PLF* | | | | 0.0127587 | 0.11 | 0.58 | 0.00247 | 98.34 |
| $\chi^2(5)$ | *OABB* | 40 | 3 | 0.2982500 | 0.2911428 | −2.38 | −2.97 | 0.23887 | — |
| | *KIM* | | | | 0.3024950 | 1.42 | 1.76 | 0.24120 | 98.02 |
| | *PLF* | | | | 0.2994254 | 0.39 | 0.48 | 0.24556 | 98.54 |
| | *OABB* | 40 | 10 | 0.2644750 | 0.2578060 | −2.52 | −5.54 | 0.12051 | — |
| | *KIM* | | | | 0.2650633 | 0.22 | 0.47 | 0.12466 | 93.18 |
| | *PLF* | | | | 0.2661069 | 0.62 | 1.31 | 0.12421 | 93.86 |
| | *OABB* | 60 | 3 | 0.1882963 | 0.1840792 | −2.24 | −4.01 | 0.10531 | — |
| | *KIM* | | | | 0.1896737 | 0.73 | 1.20 | 0.11438 | 84.64 |
| | *PLF* | | | | 0.1871717 | −0.60 | −1.05 | 0.10700 | 96.72 |
| | *OABB* | 60 | 10 | 0.1731556 | 0.1703995 | −1.59 | −4.53 | 0.06092 | — |
| | *KIM* | | | | 0.1740217 | 0.50 | 1.38 | 0.06276 | 94.03 |
| | *PLF* | | | | 0.1735171 | 0.21 | 0.58 | 0.06197 | 96.44 |
| | *OABB* | 80 | 3 | 0.1331458 | 0.1321999 | −0.71 | −2.01 | 0.04706 | — |
| | *KIM* | | | | 0.1338785 | 0.55 | 1.50 | 0.04874 | 93.21 |
| | *PLF* | | | | 0.1332633 | 0.09 | 0.25 | 0.04743 | 98.41 |
| | *OABB* | 80 | 10 | 0.1274437 | 0.1263446 | −0.86 | −3.27 | 0.03362 | — |
| | *KIM* | | | | 0.1273322 | −0.09 | −0.32 | 0.03432 | 95.87 |
| | *PLF* | | | | 0.1274068 | −0.03 | −0.11 | 0.03389 | 98.34 |

*TV*, *AE*, *RB*, *RMSE*, and *REL.EFF* stand for the true value, average estimate, relative bias, root mean-square error, and relative efficiency, respectively. $r$ and $m$ correspond to the number of observed values and the number of imputations, respectively. *OABB* stands for original approximate Bayesian bootstrap, and *KIM* and *PLF* represent the approach taken by Kim (2002) and Parzen et al. (2005), respectively. (The latter comes from the initials of the three authors.)

of the parameter, multiplied by 100, $100 \times \frac{E(\hat{V})-V}{V}$. The term "relative bias" (*RB*) used by Kim (2002) and Parzen et al. (2005) matches the *SB* and *PB*, respectively. For the purpose of congeniality with these works, we considered both accuracy measures. Below, *RB*1 and *RB*2 stand for *SB* and *PB*, respectively. Arguably, the root-mean-square error (*RMSE*) is regarded as the best criterion for evaluating $\hat{V}$ in terms of combined accuracy and precision. $RMSE(\hat{V})$ is defined as $\sqrt{E_V[(\hat{V} - V)^2]}$. Finally, the relative efficiency (*REL.EFF*) represents the ratio of variability of the variance estimates across 10,000 replications in the *ABB* and in the proposed bias–reduction techniques.

The results from small- and moderate-sample scenarios were tabulated in Tables 1 and 2, respectively. In Tables 1 and 2, *TV*, *AE*, *RB*, *RMSE*, and *REL.EFF* stand for true value of $V$, average estimate, relative bias, root mean-square error, and relative efficiency (expressed in terms of percentages), respectively. *OABB* is the abbreviated version of the original *ABB*. Finally, *KIM* and *PLF* represent the approach taken by Kim (2002) and Parzen et al. (2005), respectively. (The latter comes from the initials of the three authors.)

When the sample size is small ($n = 20$, Table 1), *PLF* and *KIM* perform better than *OABB* in terms of bias. However, efficiency losses can be substantial. Here, we find that *KIM* is between 49.06% and 86.10% efficient, and *PLF* is between 71.03% and 92.14% efficient compared to *OABB*—improved accuracy does not come for free. Given the opposing bias and efficiency trends, it is natural to examine *RMSE*s. From an *RMSE* standpoint, the performance of *OABB* turns out to be uniformly superior to both *KIM* and *PLF*, for all the scenarios in Table 1, with *PLF* better than *KIM* (in parallel with the magnitude of biases and the level of efficiency losses).

For moderate samples ($n = 100$, Table 2), the comparative efficiency of *KIM* and *PLF* are as low as 84.64% and 93.86%, respectively. *RMSE*s are the lowest under *OABB*, again for all scenarios, with mixed performances between *KIM* and *PLF*. Expectedly, with larger sample sizes, efficiency differences between *OABB* and the other two become less substantial. However, when the efficiency loss is small, *OABB* is not heavily biased anyway. It is important to note that *RMSE* and *REL.EFF* move in the same direction in both tables, suggesting that the variability is a more dominant component of *RMSE* in comparison to the bias in this particular context, further raising a warning flag in regard to assessing the true behavior of the suggested refinements on the bias of the variance estimators in the *ABB*. For the purpose of brevity, we do not include comments as to how inferences change with respect to the different levels of $m$, $r$ and the underlying distribution of complete data since they are inconsequential for the point of this note.

We do not take a fully negative position for these methods, and it is evident that they are doing a decent job for bias reduction. However, when *REL.EFF* and *RMSE* are considered, the price to be paid for better accuracy becomes apparent. Our limited simulations appear to support the conclusion that bias improvements are outweighed by efficiency losses for the variance estimator to varying degrees. The original *ABB* seems to have better small-sample properties than the modified versions in terms of the integrated behavior of accuracy and precision.

The scope of this note is limited to simple scenarios with univariate data. Despite the unsophisticated nature of our simulation study (which is completely identical to the one that appeared in the papers we comment on) and the potential generalizability issues, it is critical to evaluate the suggested improvements on a broader basis that addresses efficiency considerations as well as biases. From what we have found, we warn readers that the proposed modifications may be inferior to the original *ABB* depending on the primary focus of interest, hence should be used with caution.

## References

Kim, J.K., 2002. A note on approximate Bayesian bootstrap imputation. Biometrika 89, 470–477.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. second ed. Wiley, New York.

Parzen, M., Lipsitz, S.R., Fitzmaurice, G.M., 2005. A note on reducing the bias of the approximate Bayesian bootstrap imputation variance estimator. Biometrika 92, 971–974.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Rubin, D.B., 1996. Multiple imputation after 18+ years (with discussion). J. Amer. Statist. Assoc. 91, 473–520.

Rubin, D.B., Schenker, N., 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Amer. Statist. Assoc. 81, 366–374.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

Schafer, J.L., 1999. Multiple imputation: a primer. Statist. Methods Med. Res. 8, 3–15.