Statistical Science
2021, Vol. 36, No. 3, 400–420
https://doi.org/10.1214/20-STS793

Maximum Likelihood Multiple Imputation: Faster Imputations and Consistent Standard Errors Without Posterior Draws

Paul T. von Hippel and Jonathan W. Bartlett

Abstract. Multiple imputation (MI) is a method for repairing and analyzing data with missing values. MI replaces missing values with a sample of random values drawn from an *imputation model*. The most popular form of MI, which we call *posterior draw multiple imputation* (PDMI), draws the parameters of the imputation model from a Bayesian posterior distribution. An alternative, which we call *maximum likelihood multiple imputation* (MLMI), estimates the parameters of the imputation model using maximum likelihood (or equivalent). Compared to PDMI, MLMI is faster and yields slightly more efficient point estimates.

A past barrier to using MLMI was the difficulty of estimating the standard errors of MLMI point estimates. We derive, implement and evaluate three consistent standard error formulas: (1) one combines variances within and between the imputed datasets, (2) one uses the score function and (3) one uses the bootstrap with two imputations of each bootstrapped sample. Formula (1) modifies for MLMI a formula that has long been used under PDMI, while formulas (2) and (3) can be used without modification under either PDMI or MLMI. We have implemented MLMI and the standard error estimators in the *mlmi* and *bootImpute* packages for R.

Key words and phrases: Missing data, incomplete data.

1. INTRODUCTION

Multiple imputation (MI) is a popular method for repairing and analyzing data with missing values [27]. Under MI, the distribution of missing values is assumed to depend on the observed values Y_{obs} and an imputation model with parameter vector θ. Then MI proceeds in two steps:

1. Obtain a parameter estimate $\widehat{\theta}_{\text{obs},m}$ from Y_{obs} alone.
2. Fill in each missing value with a random imputation drawn conditionally on Y_{obs} and $\widehat{\theta}_{\text{obs},m}$.

These steps iterate multiple times ($m = 1, \ldots, M$), returning M imputed copies of the dataset. These MI data are analyzed to produce an MI point estimate $\widehat{\theta}_{\text{MI}}$ and an estimate of its variance $V_{\text{MI}} = V(\widehat{\theta}_{\text{MI}})$. Under some circumstances, discussed later [7], MI data can also be analyzed to estimate additional quantities that are not the same as the imputation parameters θ.

Different estimates $\widehat{\theta}_{\text{obs},m}$ can be used for θ. The most common approach draws $\widehat{\theta}_{\text{obs},m}$ at random from the Bayesian posterior distribution of the parameters given Y_{obs} [27]. We call estimates drawn in this way *posterior draws* (PD), or $\widehat{\theta}_{\text{PD},m}$, and when PD estimates are used in the imputation model, we call the approach *posterior draw multiple imputation*.

An alternative is to estimate the imputation parameters by applying *maximum likelihood* (ML) to the incomplete data Y_{obs} [21, 24, 33, 34, 37]. Imputation parameters estimated in this way are ML estimates, $\widehat{\theta}_{\text{ML}}$, and when ML estimates are used in the imputation model, we call the approach *maximum likelihood multiple imputation* (MLMI). Any approach that uses asymptotically efficient estimates of the imputation parameters is equivalent to MLMI.

Although PDMI is by far the more common approach in practice, it does have certain disadvantages. A minor disadvantage is that PDMI point estimates are less efficient than MLMI point estimates, but the difference in efficiency is trivial unless the fraction of missing information is large and the number of imputations M is very

Paul T. von Hippel is Associate Professor, LBJ School of Public Affairs, University of Texas, Austin, Texas, USA 78712, (e-mail: paulvonhippel@utexas.edu). Jonathan W. Bartlett is Reader in Statistics, University of Bath, BA2 7AY, UK (e-mail: j.w.bartlett@bath.ac.uk).

small [37]. In small samples, PDMI point estimates can also have more bias than MLMI point estimates, but the biases are trivial in moderate to large samples [33, 34].

The more serious disadvantage of PDMI is computational. PDMI software users sometimes report runtimes or hours or even days in large datasets [11, 18, 23, 25]. Although increases in computing power should have speeded PDMI up, in practice these increases have been offset by growth in the size of datasets and growth in the recommended number of imputations M. In early MI research, $M = 3 - 10$ imputations were recommended as adequate for stable point estimates [27], but more recent research, evaluating the stability of standard error estimates and confidence intervals, calls for as many as $M = 20 - 200$ imputations in data with a high fraction of missing information [35].

In addition to long runtimes, PDMI software can be "fussy," sometimes failing to converge [16], giving errors and warning messages that seem inscrutable to end users [25], or requiring diagnostics and changes to the prior distribution that few end users, or even experts, are qualified to carry out [16, 19, 28]. Long runtimes and convergence issues contribute to the impression—not uncommon among applied researchers—that MI is not worth the trouble. This limits the adoption of MI, which is still rare in some applied fields, such as economics.

Most of these problems occur because most PDMI software uses a computationally intensive Markov Chain Monte Carlo (MCMC) algorithm known as *data augmentation* [28]. Faster algorithms are available to get PD estimates [22], and estimation can be further accelerated by running the algorithm in parallel [17, 29]. But efforts to speed up PD estimates beg the question of whether we need PD estimates at all.

Cannot we do imputation without posterior draws, as MLMI does? No matter what we do to speed up PDMI, MLMI will always be faster, and MLMI point estimates will always be more efficient. Why then has not MLMI been used more often?

A major barrier to MLMI's adoption has been a lack of convenient formulas for estimating the variance of MLMI point estimates. The variance of PDMI point estimates can be estimated by a simple *within-between* (WB) formula (5.5) that combines variances within and between the imputed datasets [27]. But that WB formula, when applied to MLMI data, will produce variance estimates that are too small on average. For that reason, MLMI has been labeled "improper" [27], and perhaps that label has discouraged investigation. Alternative formulas have been proposed for variance estimation under MLMI [24, 37], but the formulas are cumbersome and require statistical quantities that are often unavailable in applied data analysis.

In this article, we make MLMI more usable by deriving three simpler estimators for the variance of MLMI point estimates. One formula (5.16) modifies the WB formula that is used with PDMI. One formula (6.14) simplifies a score-based (SB) variance formula first proposed by Wang and Robins [37]. And one formula (8.4) combines MI with the bootstrap to calculate variance components due to sampling and imputation. We have implemented these estimators in the *mlmi* and *bootImpute* packages for R, which we have published on the Comprehensive R Archive Network (CRAN) [4, 5].

With these new variance formulas, MLMI becomes a more practical alternative to PDMI. The rest of this article derives the variance estimators, compares their properties analytically and through simulation, and demonstrates their use in an applied data analysis.

2. INCOMPLETE DATA

Before describing different estimators, let us define the missing data problem.

If we had complete data Y_{com} with N cases, we could maximize its likelihood to get a complete-data ML estimate $\widehat{\theta}_{\text{com}}$ of the parameter vector θ. But instead we have incomplete data where some values Y_{mis} are missing and other values Y_{obs} are observed. If values are missing at random (MAR)—so that the probability of a value being missing depends only on Y_{obs}—then we can get a consistent ML estimate $\widehat{\theta}_{\text{ML}}$ using only Y_{obs}, without modeling the process that causes values to be missing [26]. Note that $\widehat{\theta}_{\text{ML}}$ is calculated from *all* the observed values, including observed values in cases with missing values [2, 9].

The variance $V_{\text{ML}} = V(\widehat{\theta}_{\text{ML}})$ of the observed-data ML estimate exceeds the variance $V_{\text{com}} = V(\widehat{\theta}_{\text{com}})$ that we would get if we had complete data. So the information V_{ML}^{-1} in in the observed data is less than the information V_{com}^{-1} that the complete data would provide. The difference is the *missing information*:

$$V_{\text{mis}}^{-1} = V_{\text{com}}^{-1} - V_{\text{ML}}^{-1}. \tag{2.1}$$

The ratio of observed to complete information is the *fraction of observed information* γ_{obs}, and the ratio of missing to complete information is the *fraction of missing information* γ_{mis}:

$$\gamma_{\text{obs}} = V_{\text{ML}}^{-1} V_{\text{com}}, \tag{2.2}$$

$$\gamma_{\text{mis}} = V_{\text{mis}}^{-1} V_{\text{com}} = I - \gamma_{\text{obs}}. \tag{2.3}$$

If θ is a scalar, then these variances and fractions are scalars. If θ is a vector, then these "variances" are covariance matrices, and the fractions of observed and missing information are matrices as well.

3. MULTIPLE IMPUTATION

MI is an algorithm with M iterations. In iteration $m = 1, \ldots, M$, MI carries out the following steps:

1. From the observed data Y_{obs}, obtain an observed-data estimate $\widehat{\theta}_{\text{obs},m}$.

2. Fill in the missing data Y_{mis} with random imputations $Y_{\text{imp},m}$ drawn conditionally on Y_{obs} and $\widehat{\theta}_{\text{obs},m}$. The result is a singly imputed (SI) dataset $Y_{\text{SI},m} = \{Y_{\text{obs}}, Y_{\text{imp},m}\}$.

Together, the M SI datasets make up a MI dataset Y_{MI}.

The difference between MLMI and PDMI lies in the definition of the observed-data estimator $\widehat{\theta}_{\text{obs},m}$ in step 1:

- Under MLMI, $\widehat{\theta}_{\text{obs},m}$ is the ML estimate $\widehat{\theta}_{\text{ML}}$, or another estimate that just as efficient in large samples.
- Under PDMI, $\widehat{\theta}_{\text{obs},m}$ is a PD estimate $\widehat{\theta}_{\text{PD},m}$ drawn at random from the posterior distribution of θ given Y_{obs}.

3.1 Computational Efficiency of MLMI over PDMI

The main advantage of MLMI is its computational efficiency. Under PDMI, a new PD estimate $\widehat{\theta}_{\text{PD},m}$ must be drawn in every iteration m, so both steps of the algorithm must be iterated. Under MLMI, by contrast, the observed-data ML estimate $\widehat{\theta}_{\text{ML}}$ is the same in every iteration, so we can run step 1 just once and only iterate step 2. Not iterating step 1 gives MLMI a speed advantage that increases with the number of iterations M.

Even when M is small, MLMI remains faster because it is faster to get ML estimates than it is to get PD estimates. In some simple settings (such as our simulation, later), both ML and PD estimates can be calculated using closed-form formulas; PD requires an extra step, but the extra runtime is trivial. In general settings, though, both ML and PD estimates require iterative, numerical methods, which are much more computationally intensive for PD than for ML. To get ML estimates, software can use *full information maximum likelihood* or the EM algorithm [12]. But to get PD estimates, most PDMI software uses data augmentation [28], in which the EM algorithm is only the first step. Data augmentation typically begins by using the EM algorithm to find the posterior mode of the parameters of the imputation model. It then takes a random walk around the posterior by iteratively reimputing the data and reestimating the imputation parameters from the imputed data. The reestimated parameters are PD estimates.

The main reason why data augmentation is slow to return results is that it discards results from the vast majority of iterations. It discards ("burns in"), say, the first 100 iterations to ensure that the PD estimates have converged to their posterior distribution; then it discards, say, 99 out out every 100 PD estimates, to ensure that the PD estimates are approximately uncorrelated. So $100M$ iterations may be required to get M PD estimates and M imputed datasets.

A faster and stabler way to get PD estimates is to bootstrap the incomplete data and calculate a ML estimate from each bootstrapped sample [15, 22, 31]. Both data augmentation and bootstrapped ML are faster if they run in parallel [17, 29]. But both methods of getting PD estimates are slower than ML, so PDMI is slower than MLMI.

3.2 Bootstrapped MI

A variant of MI which can be useful for variance estimation is *bootstrapped MI* (BMI). BMI is an iterative procedure with two nested loops. In iteration $b = 1, \ldots, B$,

1. Take a bootstrapped sample $Y_{\text{boot},b}$ of N cases from the incomplete data.
2. Then, in iteration $d = 1, \ldots, D$, apply MI to $Y_{\text{boot},b}$. That is,

   (a) From the observed values in $Y_{\text{boot},b}$, obtain an observed-data estimate $\widehat{\theta}_{\text{obs},bd}$.

   (b) Fill in $Y_{\text{boot},b}$'s missing values with random imputations drawn conditionally on $\widehat{\theta}_{\text{obs},bd}$ and the observed values in $Y_{\text{boot},b}$. The result is a single bootstrapped-then-imputed (BSI) dataset $Y_{\text{BSI},bd}$.

Together, the BD BSI datasets make up an BMI dataset Y_{BMI}.

There are two flavors of BMI: boostrapped MLMI (BMLMI) and bootstrapped PDMI (BPDMI). The difference is the definition of the estimator $\widehat{\theta}_{\text{obs},bd}$:

- Under BMLMI, $\widehat{\theta}_{\text{obs},bd}$ is an ML estimate $\widehat{\theta}_{\text{ML},b}$ derived from the observed values in $Y_{\text{boot},b}$.
- Under BPDMI, $\widehat{\theta}_{\text{obs},bd}$ is a PD estimate $\widehat{\theta}_{\text{PD},bd}$ drawn at random from the posterior distribution of θ given the observed values in $Y_{\text{boot},b}$.

As in other applications of the bootstrap, $B = 40$ is adequate for some purposes, though larger B is better. The optimal value for D, however, is 2, for reasons we will discuss when we get to variance estimation.

Just as MLMI is faster than PDMI, BMLMI is faster than BPDMI. Not only is $\widehat{\theta}_{\text{ML},b}$ easier to calculate than $\widehat{\theta}_{\text{PD},bd}$, but $\widehat{\theta}_{\text{ML},b}$ only needs to be calculated once for each bootstrapped sample, while $\widehat{\theta}_{\text{PD},bd}$ needs to be calculated D times for each bootstrapped sample. That is, in the bth bootstrapped sample, PDMI must iterate all of step 2, while MLMI can run step 2(a) just once and only iterate step 2(b).

4. MI POINT ESTIMATES

With large N, and large M or BD, practically equivalent point estimates can be calculated from data that was imputed using PDMI or MLMI, with or without the bootstrap. With modest M or BD, however, MLMI point estimates are more efficient than PDMI point estimates, and point estimates from either MLMI or PDMI are more efficient without the bootstrap than with it. This section shows why.

There are several ways to get point estimates from MI data. The most common way is *repeated MI* [27], which analyzes each SI dataset as though it were complete, producing M SI point estimates $\widehat{\theta}_{\mathrm{SI},m}$, $m = 1, \ldots, M$, whose average is a repeated MI point estimate:

$$\widehat{\theta}_{\mathrm{MI}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\theta}_{\mathrm{SI},m}. \tag{4.1}$$

Under MLMI, we call this estimate $\widehat{\theta}_{\mathrm{MLMI}}$; under PDMI we call it $\widehat{\theta}_{\mathrm{PDMI}}$. The corresponding SI estimators are $\widehat{\theta}_{\mathrm{MLSI}}$ and $\widehat{\theta}_{\mathrm{PDSI}}$. The limit of $\widehat{\theta}_{\mathrm{MI}}$ as M gets large is $\lim_{M\to\infty} \widehat{\theta}_{\mathrm{MI}} = \theta_{\infty I}$.

We can also get point estimates from BMI data. Analyze each of the bootstrapped-then-imputed datasets as though it were complete to obtain *BD* individual point estimates $\widehat{\theta}_{\mathrm{BD}}$. Then average the individual estimates $\widehat{\theta}_{\mathrm{BD}}$ to get a BMI point estimate:

$$\widehat{\theta}_{\mathrm{BMI}} = \frac{1}{\mathrm{BD}}\sum_{b=1}^{B}\sum_{d=1}^{D} \widehat{\theta}_{bd}. \tag{4.2}$$

Under BMLMI, we call this estimate $\widehat{\theta}_{\mathrm{BMLMI}}$; under BPDMI, we call it $\widehat{\theta}_{\mathrm{BPDMI}}$.

4.1 Variance of MI Point Estimates

Repeated MI point estimates $\widehat{\theta}_{\mathrm{MI}}$ are consistent, asymptotically normal and approach $\widehat{\theta}_{\mathrm{ML}}$ as M and N get large. The variance of a MI point estimate is

$$V_{\mathrm{MI}} = V(\widehat{\theta}_{\mathrm{MI}}) = V_{\mathrm{ML}} + \frac{1}{M}(V_{\mathrm{SI}} - V_{\mathrm{ML}}), \tag{4.3}$$

where V_{SI} is the variance of an SI point estimates. So as M gets large, the variance of an MI point estimate approaches V_{ML}. All these statements are true under both MLMI and PDMI [37].

When M is finite, though, MLMI point estimates have smaller variance than PDMI point estimates. To see why, notice that the variance of an MI point estimate depends in part on the variance of the observed-data estimate $\widehat{\theta}_{\mathrm{obs},m}$ that is used to generate imputations—and in large samples the most efficient possible observed-data estimate is a ML estimate. In fact, PD estimates are approximately twice as variable as ML estimates [33, 34]. To see why, notice that $\widehat{\theta}_{\mathrm{PD},m}$ is drawn from a posterior density whose asymptotic distribution is $\widehat{\theta}_{\mathrm{PD},m} \sim N(\widehat{\theta}_{\mathrm{ML}}, \widehat{V}_{\mathrm{ML}})$. So the variance of $\widehat{\theta}_{\mathrm{PD},m}$ is $V_{\mathrm{PD}} = V(\widehat{\theta}_{\mathrm{ML}}) + \widehat{V}_{\mathrm{ML}} \approx 2V_{\mathrm{ML}}$.

The substantial efficiency advantage of $\widehat{\theta}_{\mathrm{ML}}$ over $\widehat{\theta}_{\mathrm{PD}}$ translates into a smaller efficiency advantage of $\widehat{\theta}_{\mathrm{MLMI}}$ over $\widehat{\theta}_{\mathrm{PDMI}}$. With large N, the variances of $\widehat{\theta}_{\mathrm{MLMI}}$ and $\widehat{\theta}_{\mathrm{PDMI}}$ are

$$V_{\mathrm{MLMI}} = V(\widehat{\theta}_{\mathrm{MLMI}}) = V_{\mathrm{ML}} + \frac{1}{M}V_{\mathrm{com}}\gamma_{\mathrm{mis}}, \tag{4.4}$$

$$V_{\mathrm{PDMI}} = V(\widehat{\theta}_{\mathrm{PDMI}}) = V_{\mathrm{ML}} + \frac{1}{M}V_{\mathrm{ML}}\gamma_{\mathrm{mis}}. \tag{4.5}$$

These expressions come from Wang and Robins ([37], equations (1) and (2)), but we have simplified the expression for V_{PDMI}; the steps of the simplification are given in Appendix A.

Since $V_{\mathrm{com}} < V_{\mathrm{ML}}$ it follows that $V_{\mathrm{MLMI}} < V_{\mathrm{PDMI}}$—that is, MLMI is more efficient than PDMI in large samples. In small samples, MLMI is also more efficient than PDMI, and can be less biased as well, at least in normal data [33, 34].

Later it will be helpful to have expressions for the variance of the SI estimators. We can get those expressions by taking the variance of the MI estimators and setting $M = 1$:

$$V_{\mathrm{MLSI}} = V(\widehat{\theta}_{\mathrm{MLSI}}) = V_{\mathrm{ML}} + V_{\mathrm{com}}\gamma_{\mathrm{mis}}, \tag{4.6}$$

$$V_{\mathrm{PDSI}} = V(\widehat{\theta}_{\mathrm{PDSI}}) = V_{\mathrm{ML}} + V_{\mathrm{ML}}\gamma_{\mathrm{mis}}. \tag{4.7}$$

4.2 Variance of BMI Point Estimates

The variance of BMI point estimates is a little different. It can be calculated as follows. In large samples, the individual bootstrapped-then-imputed point estimates $\widehat{\theta}_{bd}$ fit a random effects model that is centered around $\widehat{\theta}_{\mathrm{ML}}$:

$$\widehat{\theta}_{bd} = \widehat{\theta}_{\mathrm{ML}} + e_b + e_{bd}, \tag{4.8}$$

where e_b represents bootstrap or sampling variation, and e_{bd} represents imputation variation. The variance components are

$$V(e_b) = V_{\mathrm{ML}} \tag{4.9}$$

and

$$\begin{aligned} V(e_{bd}) &= V_{\mathrm{SI}} - V_{\mathrm{ML}} \\ &= \begin{cases} V_{\mathrm{com}}\gamma_{\mathrm{mis}} & \text{under BMLMI,} \\ V_{\mathrm{ML}}\gamma_{\mathrm{mis}} & \text{under BPDMI,} \end{cases} \end{aligned} \tag{4.10}$$

where the expressions in the final brace come from substituting (4.6) and (4.7) for V_{SI}.

The BMI point estimate is just the average $\frac{1}{\mathrm{BD}}\sum\sum\widehat{\theta}_{bd}$, so its variance is

$$V_{\mathrm{BMI}} = V(\widehat{\theta}_{\mathrm{BMI}}) = V_{\mathrm{ML}} + \frac{V_{\mathrm{ML}}}{B} + \frac{V_{\mathrm{SI}} - V_{\mathrm{ML}}}{\mathrm{BD}}. \tag{4.11}$$

Clearly, V_{BMI} decreases faster with B than with D, so it makes sense to set D as low as possible. We recommend $D = 2$ since at least 2 imputations per bootstrap sample are needed for variance estimation.

With B bootstrap samples each imputed D times, a $\widehat{\theta}_{\mathrm{BMI}}$ point estimate is more variable than a nonbootstrapped MI point estimate $\widehat{\theta}_{\mathrm{MI}}$ with $M = \mathrm{BD}$ imputations. The difference in variance

$$V_{\mathrm{BMI}} - V_{\mathrm{MI}} = \frac{V_{\mathrm{ML}}}{B} \tag{4.12}$$

is obtained by subtracting (4.3) from (4.11) with $\mathrm{BD} = M$. Again, it is clear that V_{BMI} is smaller when B is large

and $D = M/B$, perforce, is small. That is one reason we recommend setting $D = 2$.

The variance of a BMI point estimates is smaller under BMLMI than under BPDMI. We get the following expressions by substituting (4.6) and (4.7) for V_{SI} in (4.11):

$$V_{\mathrm{BMLMI}} = V_{\mathrm{ML}}\left(1 + \frac{1}{B}\right) + \frac{1}{\mathrm{BD}} V_{\mathrm{com}}\gamma_{\mathrm{mis}} \quad \text{under BMLMI,} \tag{4.13}$$

$$V_{\mathrm{BPDMI}} = V_{\mathrm{ML}}\left(1 + \frac{1}{B}\right) + \frac{1}{\mathrm{BD}} V_{\mathrm{ML}}\gamma_{\mathrm{mis}} \quad \text{under BPDMI.} \tag{4.14}$$

Since $V_{\mathrm{com}} < V_{\mathrm{ML}}$, it follows that $V_{\mathrm{BMLMI}} < V_{\mathrm{BPDMI}}$.

4.3 How Many Imputations Are Needed for Point Estimates?

How many imputations are needed to produce MI point estimates that are almost as efficient as they would be with infinite imputations? The answer depends on the fraction of missing information γ_{mis} and on whether MLMI or PDMI is used. The large-N efficiencies of $\widehat{\theta}_{\mathrm{MLMI}}$ and $\widehat{\theta}_{\mathrm{PDMI}}$, relative to $\widehat{\theta}_{\mathrm{ML}}$, are

$$\mathrm{re}_{\mathrm{MLMI}} = V_{\mathrm{MLMI}}^{-1} V_{\mathrm{ML}} = \left(I + \frac{1}{M}\gamma_{\mathrm{obs}}\gamma_{\mathrm{mis}}\right)^{-1}, \tag{4.15}$$

$$\mathrm{re}_{\mathrm{PDMI}} = V_{\mathrm{PDMI}}^{-1} V_{\mathrm{ML}} = \left(I + \frac{1}{M}\gamma_{\mathrm{mis}}\right)^{-1}. \tag{4.16}$$

These relative efficiencies were calculated from expressions (4.4) and (4.5). The expression for $\mathrm{re}_{\mathrm{PDMI}}$, derived a different way, also appears in Rubin [27], page 114.[1]

Under BMI, the efficiencies of $\widehat{\theta}_{\mathrm{BMLMI}}$ and $\widehat{\theta}_{\mathrm{BPDMI}}$, relative to $\widehat{\theta}_{\mathrm{ML}}$, are

$$\mathrm{re}_{\mathrm{BMLMI}} = V_{\mathrm{MLMI}}^{-1} V_{\mathrm{ML}} = \left(\left(1 + \frac{1}{B}\right)I + \frac{1}{\mathrm{BD}}\gamma_{\mathrm{obs}}\gamma_{\mathrm{mis}}\right)^{-1}, \tag{4.17}$$

$$\mathrm{re}_{\mathrm{BPDMI}} = V_{\mathrm{PDMI}}^{-1} V_{\mathrm{ML}} = \left(\left(1 + \frac{1}{B}\right)I + \frac{1}{\mathrm{BD}}\gamma_{\mathrm{mis}}\right)^{-1}. \tag{4.18}$$

These efficiencies were calculated from (4.11).

Table 1 shows the number of imputations that are needed for MI point estimates to have 95% asymptotic relative efficiency. Under MI, the number of imputations is M; under BMI, it is BD with $D = 2$.

[1] Rubin was estimating the efficiency of a PDMI point estimate with M imputations relative to one with infinite imputations, whereas we are calculating the efficiency of a PDMI estimate relative to an ML estimate. In large samples, however, a ML estimate is equivalent to a PDMI estimate with infinite imputations, so the two definitions of asymptotic efficiency are the same.

TABLE 1
Number of imputations needed for point estimates to have 95% asymptotic relative efficiency

	Imputations needed			
γ_{mis}	PDMI	MLMI	BPDMI	BMLMI
0.1	2	2	38	36
0.2	4	3	38	38
0.3	6	4	40	38
0.4	8	4	42	40
0.5	10	4	44	40
0.6	12	4	46	40
0.7	14	4	48	38
0.8	16	3	50	38
0.9	18	2	50	36

Note. For PDMI and MLMI, the number of imputations shown is M. For BPDMI and BMLMI, the number of imputations shown is BD, where B is the number of bootstrap samples and $D = 2$ is the number of imputations per bootstrap sample.

MLMI point estimates need fewer imputations than PDMI point estimates, especially when γ_{mis} is large. Under PDMI, the number of imputations needed increases linearly as $M = 2\gamma_{\mathrm{mis}}$, but under MLMI, M is a quadratic function of γ_{mis} that peaks at $M = 4$ near $\gamma_{\mathrm{mis}} = 0.5$ and falls if γ_{mis} is larger or smaller. PDMI and MLMI need similar numbers of imputations if γ_{mis} is small, but if γ_{mis} is large MLMI needs many fewer imputations. For example, if $\gamma_{\mathrm{mis}} = 0.9$, MLMI needs just 2 imputations while PDMI needs 18 imputations to achieve the same efficiency.

Under BMI, BMLMI needs fewer imputations than BPDMI to achieve point estimates with the same efficiency. But the difference is relatively small. Using either form of BMI, 38 to 50 imputations typically suffice, that is,, 19 to 25 bootstrapped datasets, each imputed twice.

If the efficiency of point estimates were all that mattered, we would clearly choose MLMI over PDMI, and we would not give BMI a second thought. But the picture changes somewhat when we go beyond point estimates and consider variance estimates as well.

5. WB VARIANCE ESTIMATES

In the coming sections, we will derive three ways to estimate the variance of an MI point estimate. We call these the WB variance estimate, the SB variance estimate, and the bootstrapped MI variance estimate. Each variance estimate can be used to calculate a confidence interval (or hypothesis test) and estimate the fraction of missing information. Both WB and SB estimates make certain assumptions about the imputation and analysis model, which we will discuss later. Bootstrapped MI makes fewer assumptions.

This section derives the *within-between* (*WB*) *estimators*, so-called because they rely on variance components that lie within and between the SI datasets in MI data.

When we analyze a SI dataset as though it were complete, we get not just a SI point estimate $\widehat{\theta}_{\mathrm{SI},m}$ but also a SI variance estimate $\widehat{V}_{\mathrm{com,SI},m}$ that would consistently estimate the variance if the data were complete. Across the M SI datasets, the average of the $\widehat{V}_{\mathrm{com,SI},m}$ is the within variance $\widehat{W}_{\mathrm{MI}}$, and the variance of the SI point estimates $\widehat{\theta}_{\mathrm{SI},m}$ is the between variance $\widehat{B}_{\mathrm{MI}}$.

$$\widehat{W}_{\mathrm{MI}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{V}_{\mathrm{com,SI},m}, \tag{5.1}$$

$$\widehat{B}_{\mathrm{MI}} = \frac{1}{M-1}\sum_{m=1}^{M} (\widehat{\theta}_{\mathrm{SI},m} - \widehat{\theta}_{\mathrm{MI}})^{\otimes 2}. \tag{5.2}$$

Here, the notation $(\widehat{\theta}_{\mathrm{SI},m} - \widehat{\theta}_{\mathrm{MI}})^{\otimes 2}$ represents the outer product $(\widehat{\theta}_{\mathrm{SI},m} - \widehat{\theta}_{\mathrm{MI}})(\widehat{\theta}_{\mathrm{SI},m} - \widehat{\theta}_{\mathrm{MI}})^T$, which reduces to the square $(\widehat{\theta}_{\mathrm{SI},m} - \widehat{\theta}_{\mathrm{MI}})^2$ if θ is scalar [37].

Clearly, $\widehat{W}_{\mathrm{MI}}$ is a consistent estimator of V_{com} [27, 30]. $\widehat{B}_{\mathrm{MI}}$ is an unbiased and consistent estimator for the variance of $\widehat{\theta}_{\mathrm{SI}}$ around $\widehat{\theta}_{\infty I}$, and since $\widehat{\theta}_{\infty I}$ approaches $\widehat{\theta}_{\mathrm{ML}}$ in large samples, it follows that $\widehat{B}_{\mathrm{MI}}$ consistently estimates

$$\begin{aligned} E(\widehat{B}_{\mathrm{MI}}) &= V(\widehat{\theta}_{\mathrm{SI}} \mid \widehat{\theta}_{\infty I}) \xrightarrow{N\to\infty} V(\widehat{\theta}_{\mathrm{SI}} \mid \widehat{\theta}_{\mathrm{ML}}) \\ &= V_{\mathrm{SI}} - V_{\mathrm{ML}} \\ &= \begin{cases} V_{\mathrm{com}}\gamma_{\mathrm{mis}} & \text{under MLMI,} \\ V_{\mathrm{ML}}\gamma_{\mathrm{mis}} & \text{under PDMI.} \end{cases} \end{aligned} \tag{5.3}$$

The last line, which is obtained by substituting expressions (4.6) and (4.7) for V_{SI}, shows that $\widehat{B}_{\mathrm{MI}}$ estimates a different quantity under MLMI than under PDMI. When this distinction is important, we will use the symbols $\widehat{B}_{\mathrm{MLMI}}$ and $\widehat{B}_{\mathrm{PDMI}}$, along with $\widehat{W}_{\mathrm{MLMI}}$ and $\widehat{W}_{\mathrm{PDMI}}$.

A useful corollary of (5.3) is that $\widehat{B}_{\mathrm{MI}}/M$ is a consistent estimator for the variance of $\widehat{\theta}_{\mathrm{MI}}$ around $\widehat{\theta}_{\mathrm{ML}}$:

$$\begin{aligned} E\left(\frac{1}{M}\widehat{B}_{\mathrm{MI}}\right) &\xrightarrow{N\to\infty} \frac{1}{M} V(\widehat{\theta}_{\mathrm{SI}} \mid \widehat{\theta}_{\mathrm{ML}}) \\ &= V(\widehat{\theta}_{\mathrm{MI}} \mid \widehat{\theta}_{\mathrm{ML}}) \\ &= V_{\mathrm{MI}} - V_{\mathrm{ML}}. \end{aligned} \tag{5.4}$$

So if we derive a consistent estimator of V_{ML}, we can add $\widehat{B}_{\mathrm{MI}}/M$ to get a consistent estimator of V_{MI}.

Although consistent, $\widehat{B}_{\mathrm{MI}}$ can be imprecise when M is small, because $\widehat{B}_{\mathrm{MI}}$ is a variance estimated from a sample of just M imputations. Estimators that give substantial weight to $\widehat{B}_{\mathrm{MI}}$ will be imprecise as well. We will return to this issue repeatedly in the next couple of pages.

5.1 WB Variance Estimation Under PDMI

Under PDMI, the WB variance estimator is

$$\widehat{V}_{\mathrm{PDMI,WB}} = \widehat{W}_{\mathrm{PDMI}} + \widehat{B}_{\mathrm{PDMI}} + \frac{1}{M}\widehat{B}_{\mathrm{PDMI}}. \tag{5.5}$$

This estimator can be derived in a Bayesian framework [27], but it can also be derived by substituting consistent estimators for the components of V_{PDMI} in equation (4.5) [37]. That is, $\widehat{V}_{\mathrm{PDMI,WB}}$ consistently estimates V_{PDMI} because $\widehat{W}_{\mathrm{PDMI}}$ consistently estimates V_{com}, $\widehat{B}_{\mathrm{PDMI}}$ consistently estimates $V_{\mathrm{ML}} - V_{\mathrm{com}}$, and $\widehat{B}_{\mathrm{PDMI}}/M$ consistently estimates $V_{\mathrm{PDMI}} - V_{\mathrm{ML}}$.

These are WB estimators for the fractions of observed and missing information under PDMI:

$$\widehat{\gamma}_{\mathrm{obs|PDMI,WB}} = (\widehat{W}_{\mathrm{PDMI}} + \widehat{B}_{\mathrm{PDMI}})^{-1}\widehat{W}_{\mathrm{PDMI}}, \tag{5.6}$$

$$\widehat{\gamma}_{\mathrm{mis|PDMI,WB}} = I - \widehat{\gamma}_{\mathrm{obs|PDMI,WB}}. \tag{5.7}$$

Again the consistency of these estimators can be verified by substitution. $\widehat{\gamma}_{\mathrm{obs,PDMI,WB}}$ is consistent for $\gamma_{\mathrm{obs}} = V_{\mathrm{ML}}^{-1}V_{\mathrm{com}}$ because $\widehat{W}_{\mathrm{PDMI}}$ is consistent for V_{com} and $\widehat{W}_{\mathrm{PDMI}} + \widehat{B}_{\mathrm{PDMI}}$ is consistent for V_{ML}. It follows that $\widehat{\gamma}_{\mathrm{mis|PDMI,WB}}$ is consistent for γ_{mis}.

(In the PDMI literature, the fraction of observed information is usually defined a little differently, as $V_{\mathrm{PDMI}}^{-1}V_{\mathrm{com}}$. Under that definition, the fractions of observed and missing information are consistently estimated by $\widetilde{\gamma}_{\mathrm{obs|PDMI,WB}} = \widehat{V}_{\mathrm{PDMI}}^{-1}\widehat{W}_{\mathrm{PDMI}}$ and $\widetilde{\gamma}_{\mathrm{mis|PDMI,WB}} = I - \widetilde{\gamma}_{\mathrm{obs|PDMI,WB}}$.)

We can construct a WB confidence interval for scalar θ:

$$\widehat{\theta}_{\mathrm{PDMI}} \pm t_{\mathrm{PDMI,WB}}\widehat{V}_{\mathrm{PDMI,WB}}^{1/2}, \tag{5.8}$$

where $t_{\mathrm{PDMI,WB}}$ is a quantile from a t distribution with $\nu_{\mathrm{PDMI,WB}}$ degrees of freedom (df). A simple df estimate is

$$\widehat{\nu}_{\mathrm{PDMI,WB}} = (M-1)\widetilde{\gamma}_{\mathrm{mis|PDMI,WB}}^{-2} \tag{5.9}$$

[27], but this estimate can be highly variable and produce values that are unrealistically large (exceeding the sample size) or unnecessarily small (less than 3). To avoid these problems, we replace $\widehat{\nu}_{\mathrm{PDMI,WB}}$ with

$$\widetilde{\nu}_{\mathrm{PDMI,WB}} = \max\left(3, \left(\widehat{\nu}_{\mathrm{PDMI,WB}}^{-1} + \widetilde{\nu}_{\mathrm{obs}}^{-1}\right)^{-1}\right) \tag{5.10}$$

which is bounded below at 3 and above at the df in the observed data, estimated by

$$\widetilde{\nu}_{\mathrm{obs}} = \nu_{\mathrm{com}}\widetilde{\gamma}_{\mathrm{obs|PDMI,WB}}\left(\frac{\nu_{\mathrm{com}}+1}{\nu_{\mathrm{com}}+3}\right), \tag{5.11}$$

where ν_{com} is the df that would be available if the data were complete, for example, $\nu_{\mathrm{com}} = N - 2$ for a simple linear regression [3, 34]. If θ is a vector, we use the same formulas but replace $\widetilde{\gamma}_{\mathrm{mis|PDMI,WB}}$ with the average of its diagonal elements [3].

The WB estimators are functions of $\widehat{B}_{\text{PDMI}}$ and give more weight to $\widehat{B}_{\text{PDMI}}$ if γ_{mis} is large. Since $\widehat{B}_{\text{MI}}$ is imprecise and volatile if M is small, it follows that the WB estimators are imprecise and volatile if M is small and γ_{mis} is large. The number of imputations M that are needed for stable variance estimates increases quadratically with γ_{mis} [35]:

$$M = 1 + \frac{1}{2}(\gamma_{\text{mis}}/\text{CV})^2, \tag{5.12}$$

where CV is the desired coefficient of variation for the SE estimate. For example, if we want CV = 0.05—implying that the SE estimate is expected to change by about 5% if we impute the data again—then we should use $M = 1 + 200\gamma_{\text{mis}}^2$ imputations, for example, just 3 imputations if $\gamma_{\text{mis}} = 0.1$ but 51 imputations if $\gamma_{\text{mis}} = 0.5$.

5.2 WB Variance Estimation Under MLMI

The WB formulas that are consistent under PDMI are inconsistent under MLMI, and for that reason MLMI has been defined as "improper." But we now present alternative WB estimators that are consistent under MLMI:

$$\widehat{\gamma}_{\text{mis|MLMI,WB}} = \widehat{W}_{\text{MLMI}}^{-1}\widehat{B}_{\text{MLMI}}, \tag{5.13}$$

$$\widehat{\gamma}_{\text{obs|MLMI,WB}} = I - \widehat{\gamma}_{\text{mis|MLMI,WB}}, \tag{5.14}$$

$$\widehat{V}_{\text{ML|MLMI,WB}} = \widehat{W}_{\text{MLMI}}\widehat{\gamma}_{\text{obs|MLMI,WB}}^{-1}, \tag{5.15}$$

$$\widehat{V}_{\text{MLMI,WB}} = \widehat{V}_{\text{ML|MLMI,WB}} + \frac{1}{M}\widehat{B}_{\text{MLMI}}. \tag{5.16}$$

To verify the consistency of these estimators, replace $\widehat{W}_{\text{MLMI}}$, $\widehat{B}_{\text{MLMI}}$ and $\widehat{B}_{\text{MLMI}}/M$ with their estimands: $\widehat{W}_{\text{MLMI}}$ consistently estimates V_{com}, $\widehat{B}_{\text{MLMI}}$ consistently estimates $V_{\text{com}}\gamma_{\text{mis}}$ (from (5.3)), and $\widehat{B}_{\text{MLMI}}/M$ consistently estimates $V_{\text{MLMI}} - V_{\text{ML}}$ (from (5.4)).

Although consistent, the WB estimators under MLMI can be imprecise if M is small and γ_{mis} is large. The imprecision comes again from $\widehat{B}_{\text{MLMI}}$. In fact, $\widehat{B}_{\text{MLMI}}$ can be so imprecise that it exceeds $\widehat{W}_{\text{MLMI}}$. If $\widehat{B}_{\text{MLMI}}$ is scalar, the fact that it can exceed $\widehat{W}_{\text{MLMI}}$ means that the estimate $\widehat{\gamma}_{\text{mis|MLMI,WB}}$ can exceed one, although the estimand γ_{mis} cannot; therefore, the estimates $\widehat{\gamma}_{\text{obs|MLMI,WB}}$ and $\widehat{V}_{\text{ML|MLMI,WB}}$ can be negative, although the corresponding estimands must be positive. If $\widehat{B}_{\text{MLMI}}$ is a matrix, the problem is that the estimated covariance matrix $\widehat{V}_{\text{ML|MLMI,WB}}$ may not be positive definite, although the true covariance matrix must be. These problems are rare if γ_{mis} is small, but more common if γ_{mis} is large and M is small. (See Appendix B.)

To increase precision and avoid negative estimates, if $\widehat{\gamma}_{\text{mis|MLMI,WB}}$ is a scalar we replace it with a shrunken estimator that is guaranteed to take values between 0 and 1:

$$\widetilde{\gamma}_{\text{mis|MLMI,WB}} = h(\widehat{\gamma}_{\text{mis|MLMI,WB}}, M - 1). \tag{5.17}$$

Here, the shrinkage function is

$$h(\widehat{\gamma}, \nu) = \frac{\nu}{2}\widehat{\gamma}\frac{\Gamma(\frac{\nu-2}{2}, \frac{\nu}{2}\widehat{\gamma})}{\Gamma(\frac{\nu}{2}, \frac{\nu}{2}\widehat{\gamma})}, \tag{5.18}$$

where $\Gamma(a, z)$ is the upper incomplete gamma function. This shrinkage function is derived in Appendix B.

If $\widehat{\gamma}$ is a matrix, the shrinkage function becomes

$$H(\widehat{\gamma}_{\text{mis|MLMI,WB}}, \nu) = Q\widetilde{\Lambda}Q^{-1}, \tag{5.19}$$

where Q is the eigenvector matrix for $\widehat{\gamma}$, and $\widetilde{\Lambda}$ is a diagonal matrix of eigenvalues, each shrunk by $h()$. This requires that all the eigenvalues are nonzero, which in turn requires that M exceeds the number of rows in $\widehat{\gamma}$.

The shrunken estimator $\widetilde{\gamma}_{\text{mis|MLMI,WB}}$ is guaranteed to have eigenvalues between 0 and 1, and the shrunken estimator $\widetilde{V}_{\text{MLMI,WB}}$ is guaranteed to be positive definite. In addition, the shrunken variance estimator $\widetilde{V}_{\text{MLMI,WB}}$ is less variable than the nonshrunken estimator $\widehat{V}_{\text{MLMI,WB}}$. There is more shrinkage if $\widehat{\gamma}_{\text{mis|MLMI,WB}}$ is large or M is small, and less shrinkage otherwise.

Shrunken estimates of γ_{obs}, V_{ML}, and V_{MLMI} can be obtained by substituting $\widetilde{\gamma}_{\text{mis|MLMI,WB}}$ for $\widehat{\gamma}_{\text{mis|MLMI,WB}}$ in equations (5.14), (5.15) and (5.16). The shrunken estimates $\widetilde{\gamma}_{\text{obs|MLMI,WB}}$, $\widetilde{V}_{\text{ML,WB}}$ and $\widetilde{V}_{\text{MLMI,WB}}$ are guaranteed to be positive definite; they are also less variable than their nonshrunken counterparts $\widehat{\gamma}_{\text{obs|MLMI,WB}}$, $\widehat{V}_{\text{ML,WB}}$ and $\widehat{V}_{\text{MLMI,WB}}$.

The cost of shrinkage is that the shrunken estimators $\widetilde{\gamma}_{\text{mis|MLMI,WB}}$, $\widetilde{V}_{\text{ML,WB}}$ and $\widetilde{V}_{\text{MLMI,WB}}$ are biased toward zero (too small on average) if γ_{mis} is large and M is small relative to γ_{mis}. Table 2 uses numerical integration (see Appendix B) to estimate the number of imputations that are needed to avoid negative bias in $\widetilde{V}_{\text{MLMI|WB}}$. Ten or fewer imputations suffice if $\gamma_{\text{mis}} \leq 0.6$, which covers most practical settings. Above $\gamma_{\text{mis}} > 0.6$, the number of imputations required by MLMI increases quickly, but may still be practical since MLMI outputs imputations more quickly than PDMI.

TABLE 2
Number of imputations needed for approximately unbiased shrunken WB estimates under MLMI

γ_{mis}	Imputations
0.1	2
0.2	2
0.3	2
0.4	3
0.5	5
0.6	10
0.7	20
0.8	60
0.9	300

If θ is scalar, we can offer a CI:

$$\widehat{\theta}_{\mathrm{MLMI}} \pm t_{\mathrm{MLMI,WB}}\widetilde{V}^{1/2}_{\mathrm{MLMI,WB}}, \tag{5.20}$$

where $t_{\mathrm{MLMI,WB}}$ is a quantile from a t distribution whose df are approximated in Appendix C:

$$\widehat{\nu}_{\mathrm{MLMI,WB}} = \frac{\widetilde{V}^2_{\mathrm{MLMI,WB}}}{\frac{\widetilde{V}^2_{\mathrm{ML,WB}}}{\widetilde{\nu}_{\mathrm{ML,WB}}} + \frac{(\frac{1}{M}\widehat{B}_{\mathrm{MLMI}})^2}{M-1}}, \tag{5.21}$$

where

$$\widetilde{\nu}_{\mathrm{ML,WB}} = (M-1)\left(\frac{\widetilde{\gamma}_{\mathrm{obs}}}{\widetilde{\gamma}_{\mathrm{mis}}}\right)^2 - 4. \tag{5.22}$$

Notice that $\widehat{\nu}_{\mathrm{MLMI,WB}}$ converges to $\widetilde{\nu}_{\mathrm{ML,WB}}$ as M gets large.

As is the case under PDMI, under MLMI the df estimate can be highly variable and it is helpful to prevent it from getting too high or too low. To accomplish this, we adapt the PDMI formula and replace $\widehat{\nu}_{\mathrm{MLMI,WB}}$ with

$$\widetilde{\nu}_{\mathrm{MLMI,WB}} = \max\bigl(3, \widehat{\nu}^{-1}_{\mathrm{MLMI,WB}} + \widetilde{\nu}^{-1}_{\mathrm{obs}}\bigr)^{-1}), \tag{5.23}$$

where $\widetilde{\nu}_{\mathrm{obs}} = \nu_{\mathrm{com}}\widetilde{\gamma}_{\mathrm{obs|MLMI,WB}}(\frac{\nu_{\mathrm{com}}+3}{\nu_{\mathrm{com}}+1})$ estimates the df in the observed data.

If θ is a vector, we use the same df formulas but replace $\widetilde{V}_{\mathrm{ML,WB}}$, $\widetilde{V}_{\mathrm{MLMI,WB}}$ and $\widehat{B}_{\mathrm{MLMI}}$ with their diagonal elements and replace $\widetilde{\gamma}_{\mathrm{obs|MLMI,WB}}$ and $\widetilde{\gamma}_{\mathrm{mis|MLMI,WB}}$ with the average of their diagonal elements.

6. SCORE-BASED (SB) VARIANCE ESTIMATION

As an alternative to WB variance estimation, Wang and Robins [37] proposed a *score-based* (SB) variance estimator, which used the *score function*, defined using the contribution of each case to the gradient of the log likelihood. Their formula was somewhat complicated, and we derive a simpler alternative, which Appendix D shows is equivalent in large samples. The same SB formulas apply under PDMI or MLMI.

The SB formulas are less often usable than the WB formulas, because the score function is often unavailable to the user. The user typically does not see the score function when they maximize the likelihood, and some common estimation techniques, such as least squares, do not maximize the likelihood explicitly, but obtain equivalent estimates by other means. In addition, the SB formula assumes independently and identically distributed (i.i.d.) observations, which the WB formulas do not assume.

Here is a derivation of our SB formula. Let $S_{\mathrm{com}} = \nabla \ln L(\theta \mid Y_{\mathrm{com}})$ be the complete-data score that would be available with complete data, and let $S_{\mathrm{obs}} = \nabla \ln L(\theta \mid Y_{\mathrm{obs}})$ be the observed-data score that is available given the observed data. Both scores have expectations of zero. The variance of the complete-data score is the complete-data information $V^{-1}_{\mathrm{com}} = V(S_{\mathrm{com}})$. The variance of the observed-data score is the observed-data information $V(S_{\mathrm{obs}}) = V^{-1}_{\mathrm{ML}}$.

In i.i.d. data, each observation makes an equally weighted contribution to the score. In complete data, the score can be expressed as the sum $S_{\mathrm{com}} = \sum_{i=1}^{N} s_{\mathrm{com},i}$, where each summand $s_{\mathrm{com},i} = \nabla \ln L(\theta \mid y_{\mathrm{com},i})$ is a function of the parameters θ and the values $y_{\mathrm{com},i}$ of the complete data in observation i. We can think of $s_{\mathrm{com},i}$ as a variable with a different value in each observation. Then $s_{\mathrm{com},i}$ has an expectation of zero and a variance of $V(s_{\mathrm{com},i}) = V^{-1}_{\mathrm{com}}N^{-1}$.

We can estimate $s_{\mathrm{com},i}$ using MI data. For observation i in SI dataset m, the estimate is

$$\widehat{s}_{\mathrm{com},i,m} = \nabla \ln L(\widehat{\theta}_{\mathrm{MI}} \mid y_{\mathrm{SI},i,m}) \tag{6.1}$$

and the variance (over i) of $\widehat{s}_{\mathrm{com},i,m}$ consistently estimates $V^{-1}_{\mathrm{com}}N^{-1}$. In addition, $\widehat{s}_{\mathrm{com},i,m}$ can be split into random effects components. One component lies between observations, and the other component lies within observations, for example, between different imputations of the same observation:

$$\widehat{s}_{\mathrm{com},i,m} = s_{\infty I,i} + d_{\mathrm{SI},m,i}. \tag{6.2}$$

The between-observation component $s_{\infty I,i}$ is the average of $\widehat{s}_{\mathrm{com},m,i}$ across the infinite population of imputations; in large samples, $s_{\infty I,i}$ is equivalent to $s_{\mathrm{obs},i} = \nabla \ln L(\theta|y_{\mathrm{obs},i})$, which is the contribution of case i to S_{obs}. The within-observation component $d_{\mathrm{SI},m,i}$ is the imputation-specific departure of $\widehat{s}_{\mathrm{com},m,i}$ from the average $s_{\infty I,i}$. The components have expectations of zero and asymptotic variances (over i) of

$$V(\widehat{s}_{\mathrm{com},m,i}) \xrightarrow{N\to\infty} \frac{1}{N}V^{-1}_{\mathrm{com}}, \tag{6.3}$$

$$V(s_{\infty I,i}) \xrightarrow{N\to\infty} \frac{1}{N}V^{-1}_{\mathrm{ML}}, \tag{6.4}$$

$$V(d_{\mathrm{SI},m,i}) \xrightarrow{N\to\infty} \frac{1}{N}V^{-1}_{\mathrm{mis}}. \tag{6.5}$$

We can estimate the variance components using MANOVA, and multiply the variance estimates by N to obtain estimators of V^{-1}_{com}, V^{-1}_{mis}, and V^{-1}_{ML}:

$$\widehat{V}^{-1}_{\mathrm{com|SB}} = \frac{\mathrm{SST}}{M} = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}\widehat{s}^{\otimes 2}_{\mathrm{com},m,i}, \tag{6.6}$$

$$\widehat{V}^{-1}_{\mathrm{mis|SB}} = \frac{\mathrm{SSW}}{M-1} = \frac{1}{M-1}\sum_{m=1}^{M}\sum_{i=1}^{N}(\widehat{s}_{\mathrm{com},m,i} - \overline{s}_{\mathrm{com},i})^{\otimes 2}, \tag{6.7}$$

$$\widehat{V}^{-1}_{\mathrm{ML|SB}} = \widehat{V}^{-1}_{\mathrm{com|SB}} - \widehat{V}^{-1}_{\mathrm{mis|SB}}, \tag{6.8}$$

where $\overline{s}_{\mathrm{com},i} = M^{-1}\sum_{m=1}^{M}\widehat{s}_{\mathrm{com},m,i}$, and *SST* and *SSW* are the total and within sums of squares. We can use these

results to derive estimators that are consistent for γ_{mis} and γ_{obs}:

$$\widehat{\gamma}_{\text{mis|SB}} = \widehat{V}_{\text{mis|SB}}^{-1}\widehat{V}_{\text{com|SB}}, \tag{6.9}$$

$$\widehat{\gamma}_{\text{obs|SB}} = I - \widehat{\gamma}_{\text{mis}}. \tag{6.10}$$

It occasionally happens that $\widehat{V}_{\text{ML|SB}}$ and $\widehat{\gamma}_{\text{obs|SB}}$ will fail to be positive definite, especially if M is small and γ_{mis} is large. This happens when some of the eigenvalues of $\widehat{\gamma}_{\text{mis|SB}}$ exceed 1. To guarantee positive definiteness, we shrink the estimators as follows:

$$\widetilde{\gamma}_{\text{mis|SB}} = H\big(\widehat{\gamma}_{\text{mis|SB}}, (M-1)N\big), \tag{6.11}$$

$$\widetilde{\gamma}_{\text{obs|SB}} = I - \widetilde{\gamma}_{\text{mis|SB}}, \tag{6.12}$$

$$\widetilde{V}_{\text{ML|SB}} = \widehat{V}_{\text{com|SB}}\widetilde{\gamma}_{\text{obs|SB}}^{-1}, \tag{6.13}$$

where the shrinkage function $H()$ was defined in (5.19).

Then an SB estimator for the variance of an MI point estimate is

$$\widetilde{V}_{\text{MI|SB}} = \widetilde{V}_{\text{ML|SB}} + \frac{1}{M}\widehat{B}_{\text{MI}}. \tag{6.14}$$

$\widetilde{V}_{\text{MI|SB}}$ consistently estimates V_{MI} because $\widetilde{V}_{\text{ML|SB}}$ consistently estimates V_{ML} and $\widehat{B}_{\text{MI}}/M$ consistently estimates $V_{\text{MI}} - V_{\text{ML}}$.

An SB CI for scalar θ is

$$\widehat{\theta}_{\text{BMI}} \pm t_{\text{SB}}\widetilde{V}_{\text{MI|SB}}^{1/2}, \tag{6.15}$$

where t_{SB} is a quantile from a t distribution with $\text{df} = \nu_{\text{SB}}$, which is the df of $\widetilde{V}_{\text{MI|SB}}$

It remains only to estimate ν_{SB}. Since $\widehat{B}_{\text{MI}}$ has $\text{df} = M-1$ and $\widetilde{V}_{\text{ML|SB}}$ may be assumed to have df no less than $\widetilde{\nu}_{\text{obs|SB}} = \nu_{\text{com}}\widetilde{\gamma}_{\text{obs|SB}}(\frac{\nu_{\text{com}}+3}{\nu_{\text{com}}+1})$, a Satterthwaite approximation for ν_{SB} is

$$\widehat{\nu}_{\text{SB}} = \frac{\widetilde{V}_{\text{MI|SB}}^2}{\frac{\widetilde{V}_{\text{ML|SB}}^2}{\widetilde{\nu}_{\text{obs|SB}}} + \frac{(\frac{1}{M}\widehat{B}_{\text{MI}})^2}{M-1}} \tag{6.16}$$

which is very close to $\widetilde{\nu}_{\text{obs|SB}}$ unless M is very small. If N and M are large then $\widehat{\nu}_{\text{SB}}$ approaches

$$\widehat{\nu}_{\text{SB}} \xrightarrow{N,M\to\infty} \begin{cases} (M-1)\left(\dfrac{M}{\gamma_{\text{obs}}\gamma_{\text{mis}}}\right)^2 \\ \quad \text{under MLMI,} \\ (M-1)\left(\dfrac{M}{\gamma_{\text{mis}}}\right)^2 \\ \quad \text{under PDMI.} \end{cases} \tag{6.17}$$

So that asymptotic degrees of freedom are larger under MLMI than under PDMI.

7. CONDITIONS FOR CONSISTENCY OF WB AND SB VARIANCE ESTIMATES

The derivations of the WB and SB variance formulas make certain assumptions. If those assumptions are not met, then the resulting variance estimates are not necessarily consistent.

7.1 Compatible and Correctly Specified Models

The WB and SB variance formulas assume that that the same model, with the same parameters θ, is used for imputation and analysis. The formulas also assume that this model is correctly specified [37]. In practice, though, the analysis model is often different from the imputation model, and one or both models may be misspecified.

When the analysis and estimation models are different, WB and SB formulas still yield consistent variance estimates if both models are "compatible" with some *common model*, and that common model is correctly specified [7]. For example, later, in the simulations, we will consider the situation where the imputation model is a linear regression of Y on X and the analysis model is a linear regression of X on Y. If both regression models have normal residuals, then both are compatible with a common model in which (X, Y) are bivariate normal.

If the imputation and analysis models are different, but compatible and correct, then the derivations of the WB and SB variance formulas are valid provided we redefine the parameter vector θ to include all the parameters of the common model, and not just the parameters of the analysis model.

How much do the extra parameters in θ matter for the variance formulas? It depends which formula you use, as we discuss next.

7.2 Which Variance Formulas Should Include All Parameters of the Common Model?

Under PDMI, the WB formula (4.5) uses only addition; it is a weighted sum of $\widehat{W}_{\text{PDMI}}$ and $\widehat{B}_{\text{PDMI}}$. As a result, the diagonal components, that is, the squared standard error estimates—in $\widehat{V}_{\text{PDMI,WB}}$ depend only on the corresponding diagonal components of $\widehat{W}_{\text{PDMI}}$ and $\widehat{B}_{\text{PDMI}}$.

This is a nice property because it means that the PDMI WB formula can be applied to any submatrix of $\widehat{W}_{\text{PDMI}}$ and $\widehat{B}_{\text{PDMI}}$ and the resulting standard error estimates will not change. In other words, you can apply the PDMI WB formula to any subset of the parameters in θ. In fact, you can apply the PDMI WB formula, in scalar form, to each component of θ, and the standard error estimates will still be the same.

Because of this property, the standard error estimates that come from the PDMI WB formula do not change when you include parameters that are not in the analysis model but are in the common model. You can safely neglect those extra parameters; you do not even have to know what they are. When using the PDMI WB formula, you can limit your attention to the parameters in the analysis model. The resulting standard errors will be consistent if the analysis and imputation models are correct and compatible.

Under MLMI, unfortunately, the WB formula (4.4) does not have the same property. It must be applied in

matrix form, and if the imputation and analysis models are not the same, it must be applied to the whole parameter vector θ of the common model—and not just selected components, such as the parameters of the analysis model. Because the MLMI WB formula involves matrix multiplication, the diagonal elements of $\widehat{V}_{\mathrm{MLMI,WB}}$ can be affected by the off-diagonal elements of $\widehat{B}_{\mathrm{MLMI,WB}}$ and $\widehat{W}_{\mathrm{MLMI,WB}}$.

The SB variance formula (6.14) has the same issue. It must be applied in matrix form, and if the imputation and analysis models are not the same, it must be applied to the whole parameter vector θ of the common model. This is because the SB variance formula uses matrix multiplication, so the off-diagonal elements of $\widehat{V}_{\mathrm{mis|SB}}$ and $\widehat{V}_{\mathrm{com|SB}}$ can affect the diagonal elements of $\widetilde{V}_{\mathrm{MI|SB}}$. We will return to this issue in the simulations.

8. BOOTSTRAP VARIANCE ESTIMATION

Unlike the WB and SB formulas, bootstrapped MI (BMI) offers consistent variance estimates and confidence intervals with nominal coverage even when the imputation and analysis models are incompatible, or even incorrect. BMI variance formulas are straightforward and do not require matrix calculations or inclusion of parameters beyond those in the analysis model. The same BMI variance formulas are consistent under BMLMI and under BPDMI.

Remember that the individual scalar point estimates $\widehat{\theta}_{bd}$ fit this random effects model (4.8):

$$\widehat{\theta}_{bd} = \widehat{\theta}_{\mathrm{ML}} + e_b + e_{bd}. \tag{8.1}$$

The variance components are $V_{\mathrm{ML}} = V(e_b)$ and $V_{\mathrm{BD}} = V(e_{bd}) = V_{\mathrm{SI}} - V_{\mathrm{ML}}$. To estimate the variance components, we fit the model using ANOVA (or MANOVA) and use mean squared formulas:

$$\widehat{V}_{\mathrm{BD|BMI}} = \mathrm{MSW}, \tag{8.2}$$

$$\widehat{V}_{\mathrm{ML|BMI}} = \frac{\mathrm{MSB} - \mathrm{MSW}}{M}, \tag{8.3}$$

where MSB is the mean square between the bootstrapped datasets, with $\mathrm{df} = B - 1$, and MSW is the mean square within the bootstrapped datasets and between the imputed datasets, with $\mathrm{df} = B(D-1)$. Then $V_{\mathrm{BMI}} = V(\widehat{\theta}_{\mathrm{BMI}})$ is estimated by

$$\widehat{V}_{\mathrm{BMI}} = \widehat{V}_{\mathrm{ML|BMI}}\left(1 + \frac{1}{B}\right) + \frac{\widehat{V}_{\mathrm{BD|BMI}}}{\mathrm{BD}}. \tag{8.4}$$

This estimate is consistent because it replaces each component of the true variance in (4.11) with a consistent estimate.

$\widehat{V}_{\mathrm{BMI}}$ can be reexpressed as a weighted sum of independent mean squares

$$\widehat{V}_{\mathrm{BMI}} = \frac{1}{M}\left(\mathrm{MSB}\left(1 + \frac{1}{B}\right) - \mathrm{MSW}\right) \tag{8.5}$$

which according to the Satterthwaite approximation has the following df:

$$\widehat{\nu}_{\mathrm{BMI}} = \frac{(\mathrm{MSB}(B+1) - \mathrm{MSW}(B))^2}{\frac{\mathrm{MSB}^2(B+1)^2}{B-1} + \frac{\mathrm{MSW}^2 B}{D-1}}. \tag{8.6}$$

If $D = 2$, as we recommended earlier, then as B gets larger, $\widehat{\nu}_{\mathrm{BMI}}$ approaches

$$\lim_{B\to\infty, D=2} \widehat{\nu}_{\mathrm{BMI}} = B\left(1 - 2\frac{\mathrm{MSB} \times \mathrm{MSW}}{\mathrm{MSB}^2 + \mathrm{MSW}^2}\right) \tag{8.7}$$

which is just a little smaller than B if the fraction of missing information is not too large.

If θ is a scalar parameter, then a confidence interval is

$$\widehat{\theta}_{\mathrm{BMI}} \pm t_{\mathrm{BMI}} \widehat{V}_{\mathrm{BMI}}^{1/2}, \tag{8.8}$$

where t_{BMI} is a quantile from a t distribution with $\mathrm{df} = \widehat{\nu}_{\mathrm{BMI}}$. Our df and CI formulas assume a scalar θ. If θ is a vector, then the same formulas apply separately to each scalar component.

Notice that BMI variance estimation does not require an estimate of the complete-data variance V_{com}. But an estimate of V_{com} is necessary to estimate the fractions of observed and missing information. To get those estimates, start with a consistent estimate $\widehat{V}_{\mathrm{com},bd}$ obtained by analyzing each of the bootstrapped-then-imputed datasets as though it were complete. The average of the $\widehat{V}_{\mathrm{com},bd}$ is a consistent estimate of V_{com}:

$$\widehat{V}_{\mathrm{com|BMI}} = \frac{1}{\mathrm{BD}} \sum_{b=1}^{B} \sum_{m=1}^{M} \widehat{V}_{\mathrm{com},bd}. \tag{8.9}$$

It follows that

$$\widehat{\gamma}_{\mathrm{obs,BMI}} = \widehat{V}_{\mathrm{ML|BMI}}^{-1} \widehat{V}_{\mathrm{com|BMI}}, \tag{8.10}$$

$$\widehat{\gamma}_{\mathrm{mis,BMI}} = I - \widehat{\gamma}_{\mathrm{obs,BMI}} \tag{8.11}$$

are consistent estimators for the fractions of observed and missing information.

8.1 How Many Imputations Are Needed for Variance Estimation?

Table 1 gave the number of imputations that were needed for relatively efficient point estimates. But more imputations may be needed to estimate variances and CIs. At a minimum, a variance estimate should be approximately unbiased if N and M are large. Most of our variance estimates will have little or no bias even if M is small. The one exception is the WB variance estimate under MLMI, and Table 2 gave the number of imputations that were needed to reduce its bias to a negligible level.

But we often want more from a variance estimate than lack of bias. We also want variance estimates to be *replicable* in the sense that approximately the same variance estimate would be obtained if the data were imputed again, or if it were bootstrapped and imputed again. And

TABLE 3
Number of imputations needed for variance estimates with specified degrees of freedom

	Score-based variance		Within-between variance		Boostrapped variance	
γ_{mis}	PDMI	MLMI	PDMI	MLMI	PDMI	MLMI
	(a) Imputations needed for df ≥ 25					
0.1	2	2	2	2	52	52
0.2	2	2	2	3	52	52
0.3	2	2	4	7	52	52
0.4	2	2	4	14	52	52
0.5	3	2	8	30	52	52
0.6	3	2	10	67	52	52
0.7	3	2	14	159	52	52
0.8	3	2	17	465	52	52
0.9	4	2	22	2350	52	52
	(b) Imputations needed for df ≥ 100					
0.1	2	2	2	3	202	202
0.2	2	2	5	8	202	202
0.3	3	3	10	21	202	202
0.4	3	3	17	48	202	202
0.5	4	3	26	105	202	202
0.6	4	3	37	235	202	202
0.7	5	3	50	568	202	202
0.8	5	2	65	1665	202	202
0.9	5	2	82	8425	202	202

Note. For MI, the number of imputations is M. For BMI, the number of imputations is BD, where B is the number of bootstrap samples and D is the number of imputations per bootstrap sample.

we want the confidence interval derived from the variance estimate to be reasonably short,

The df of the variance estimate is a useful guide to these properties. The coefficient of variation for an SE estimate is approximately $\sqrt{1/(2\text{df})}$ [35]. So at df $= 25$ an SE estimate would likely change by about 14%, and at df $= 100$ an SE estimate would likely change by about 7%, if the data were multiply imputed again—or if it were bootstrapped and imputed again under BMI.

Table 3 gives the number of imputations M, or bootstrap samples and imputations BD, that are needed for different variance estimates to have at least 25, or at least 100, degrees of freedom.

The SB variance estimates have remarkably modest needs, requiring 5 imputations or less even when the fraction of missing information is very large. Unfortunately, SB variance estimates are often unavailable in practice, since they require a score function which the analyst may not have.

The WB estimates need few imputations when the fraction of missing information is small, but require more and more imputations as the fraction of missing information grows, especially under MLMI.

The BMI variance estimators require BD $= 2(\text{df} + 1)$ imputations, regardless of the fraction of missing information. Under PDMI, BMI needs more imputations than the WB estimator even when the fraction of missing information is as large as 0.9. Under MLMI, BMI needs more imputation than the WB estimator if the fraction of missing information is less than 0.6, but BMI needs fewer imputations than the WB estimator if the fraction of missing information is 0.6 or greater. Under MLMI, therefore, if the fraction of missing information is large there is no reason to use the WB estimator when the fraction of missing information is large; instead, switch to BMI.

Remember that BMI variance estimator is consistent under circumstances when the WB and SB estimators may be inconsistent. Therefore, BMI should be preferred when there is enough time to produce the number of imputations that it requires. And more imputations can be produced more quickly using MLMI than using PDMI.

9. SOFTWARE

The second author implemented all the methods described here and published them in new R packages called *mlmi* and *bootImpute* [4, 5].

The *mlmi* package implements MLMI and PDMI versions of four different imputation models: (1) normal linear regression of one incomplete variable on one or more complete variables, (2) the multivariate normal model for data with several incomplete continuous variables, (3) the log-linear model for data with several incomplete categorical variables and (4) the general location for a "mix" of categorical and continuous variables. The general location model can be described as a multivariate normal model whose mean is conditioned on a log-linear model of the categorical variables [28]. The *mlmi* package also implements the SB formulas and WB formulas that are appropriate for data imputed using MLMI and PDMI. When using the SB formulas, the user must specify the score function.

The *bootImpute* package implements bootstrapped MI and the formulas that are used to calculate standard errors and confidence intervals from bootstrapped MI data. The *bootImpute* package can be used with any imputation function, using either MLMI or PDMI. The *bootImpute* package includes functions that integrate it with the popular *mice* package [32], which imputes missing values using a set of regression model, and the *smcfcs* package [6], which modifies the *mice* approach to ensure that the imputation and analysis models are compatible.

The second author used these R packages to carry out simulations and an applied data analysis in R. The simulation and analysis code resides in a github repository at https://github.com/jwb133/mlmiPaper. Some of the simulations were replicated independently by the first author in SAS.

10. SIMULATIONS

In this section, we use simulation to compare the properties of MLMI and PDMI, with and without the bootstrap.

10.1 Design

We simulated N rows of standard bivariate normal data (X, Y)

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right) \tag{10.1}$$

with correlation $\rho = 0.5$, means $\mu_X = \mu_Y = 0$, and variances $\sigma_X^2 = \sigma_Y^2 = 1$. The data fit a linear regression of Y on X, or of X on Y:

$$\begin{aligned} Y &= \alpha_Y + \beta_{Y.X}X + e_{Y.X} \\ &\text{where } e_{Y.X} \sim N(0, \sigma_{Y.X}^2), \end{aligned} \tag{10.2}$$

$$\begin{aligned} X &= \alpha_X + \beta_{X.Y}Y + e_{X.Y} \\ &\text{where } e_{X.Y} \sim N(0, \sigma_{X.Y}^2). \end{aligned} \tag{10.3}$$

The parameters of both regressions have the same values: $\alpha_Y = \alpha_X = 0$, $\beta_{Y.X} = \beta_{X.Y} = \rho$ and $\sigma_{Y.X}^2 = \sigma_{X.Y}^2 = 1 - \rho^2$.

We then deleted some fraction—either $p = 0.25$ or $p = 0.5$—of Y values in one of two patterns:

- *Missing completely at random (MCAR).* Each Y value has an equal probability p of being deleted.
- *Missing at random (MAR).* Y is more likely to be deleted if X is large. In particular, Y is deleted with probability $2p\Phi(X)$, where Φ is the standard normal CDF.

For a given value of p, the fraction of observed information γ_{obs} was lower under MAR than under MCAR.

We imputed missing Y values using the following imputation model:

$$\begin{aligned} Y_i &= \widehat{\alpha}_{Y.X} + \widehat{\beta}_{Y.X}X_i + e_i \\ &\text{where } e_i \sim N(0, \widehat{\sigma}_{Y.X}^2). \end{aligned} \tag{10.4}$$

The parameter estimates $\widehat{\alpha}_{Y.X}$, $\widehat{\beta}_{Y.X}$, $\widehat{\sigma}_{Y.X}^2$ were ML estimates under MLMI and PD estimates under PDMI. In this simple setting, with X complete and Y MAR or MCAR, we could get ML and PD estimates noniteratively. We got ML estimates $\widehat{\alpha}_{Y.X,\text{ML}}$, $\widehat{\beta}_{Y.X,\text{ML}}$ by OLS regression of Y on X in the n_Y cases with Y observed; then we calculated the ML estimate $\widehat{\sigma}_{Y.X,\text{ML}}^2$ by dividing the residual sum of squares by n_Y [1]. We got PD estimates by drawing from the following distributions [20]:

$$\widehat{\sigma}_{Y.X,\text{PD}}^2 \sim \frac{n_Y}{U}\widehat{\sigma}_{Y.X,\text{ML}}^2, \tag{10.5}$$

$$\begin{bmatrix} \widehat{\alpha}_{Y.X,\text{PD}} \\ \widehat{\beta}_{Y.X,\text{PD}} \end{bmatrix} \sim N_2\left(\begin{bmatrix} \widehat{\alpha}_{Y.X,\text{ML}} \\ \widehat{\beta}_{Y.X,\text{ML}} \end{bmatrix}, \widehat{V}_{\text{ML}}\frac{\widehat{\sigma}_{Y.X,\text{PD}}^2}{\widehat{\sigma}_{Y.X,\text{ML}}^2}\right), \tag{10.6}$$

where $N_2()$ is the bivariate normal distribution, $\widehat{V}_{\text{ML}}$ is the estimated variance of the ML estimates $\widehat{\alpha}_{Y.X,\text{ML}}$, $\widehat{\beta}_{Y.X,\text{ML}}$ and U is a chi-squared random variable with degrees of freedom $n_Y - 2 + \nu_{\text{prior}}$. Here, ν_{prior} is the prior degrees of freedom, which we set conventionally to 0, although 2 is a better choice [20, 33].

In the imputed data, we regressed the incomplete variable Y on the complete variable X, and then reversed the regression, regressing X on Y. Using formulas derived in previous sections, we calculated regression point estimates and their estimated covariance matrix, along with standard error estimates and confidence intervals.

When using matrix formulas to calculate the covariance matrix of the regression estimates the question arose how large a matrix we must use. As discussed in Section 7.2, the answer depends on whether the imputation model and the analysis model were the same:

- When the analysis regressed Y on X, the analysis model was the same as the imputation model, and we could limit calculations to the 2×2 covariance matrix of the parameter estimates $(\widehat{\alpha}_{Y.X}, \widehat{\beta}_{Y.X})$. (We could have used a 3×3 matrix that included $\widehat{\sigma^2}_{Y.X}$, but this was not necessary because $\widehat{\sigma^2}_{Y.X}$ is uncorrelated with $(\widehat{\alpha}_{Y.X}, \widehat{\beta}_{Y.X})$.)
- When the analysis regressed X on Y, the analysis model differed from the imputation model, but both were compatible with a common bivariate normal model for (X, Y). So the matrix calculations must use a 5×5 matrix that includes the covariances among the 5 estimated parameters of the bivariate normal distribution. There are several ways to parameterize the bivariate normal distribution. We chose the parameterization $(\alpha_{X.Y}, \beta_{X.Y}, \sigma_{X.Y}^2, \mu_Y, \sigma_Y^2)$ because it includes the parameters for the regression of X on Y.[2]

As noted in Section 7.2, the size of the covariance matrix matters only for the SB formula and the MLMI WB formula. When using the PDMI WB formula or the bootstrap formula, the size of the matrix does not affect estimated standard errors or confidence intervals.

We ran the simulation at two different sample sizes: $N = 100$ and 500. At each sample size, we used $M = 10$, 50 or 200 imputations. When using the bootstrap, we set $\text{BD} = 50$ or 200, where $B = 25$ or 100 is the number of bootstrap samples, and $D = 2$ is the number of imputations per bootstrap sample.[3] We replicated each simulated

[2] This parameterization results from factoring the bivariate normal distribution as $N_2(X, Y) = f(Y)f(X|Y)$.

[3] We considered a condition with $M = 5$ imputations, but decided against it since some B_{MI} matrices are 5×5 and would not be positive definite with $M = 5$. We also decided against a condition $B = 5$ bootstrap samples, as the resulting variance estimates would have only about 4 degrees of freedom.

condition 10,000 times, so that the coverage of 95% confidence intervals was estimated within a standard error of 0.2%.

10.2 Results

In presenting simulation results, we focus on the regression slope $\beta_{Y.X}$ or $\beta_{X.Y}$, though we got similar results, not shown, for the intercept. We summarized the accuracy of point estimates using the percent root mean squared error (RMSE), that is, the RMSE of a scalar parameter estimate $\widehat{\beta}_{Y.X}$ or $\widehat{\beta}_{X.Y}$ expressed as a percentage of the true parameter value $\beta_{Y.X}$ or $\beta_{X.Y}$. In the regression of the incomplete Y on the complete X, the estimate $\widehat{\beta}_{Y.X}$ is unbiased, so the RMSE reflects variability only. In the regression of X on Y, though, the estimate $\widehat{\beta}_{Y.X}$ is biased in small samples [34], so the RMSE reflects bias as well as variability.

10.2.1 *Regression of Y on X.* We first regressed Y on X. Here, the analysis model is the same as the imputation model, so all matrix calculations are limited to the two model parameters $(\alpha_{Y.X}, \beta_{Y.X})$. See Section 10.1 for explanation.

Table 4(a) gives the percent RMSE for point estimates of the slope $\beta_{Y.X}$. The RMSE is slightly smaller under repeated MI than under bootstrapped MI, and slightly smaller under MLMI than under PDMI. But most differences in RMSE are very small, even when there is little information or few imputations. For example, even with 10 imputations and 50% of values MAR, the RMSE is only 2% smaller under MLMI than under PDMI.

Table 4(b) gives the mean length of nominal 95% CIs, along with their departure from 95% coverage. Bootstrapped and SB CIs come within 0.5% of nominal cov-

TABLE 4
Estimating the slope of Y on X

(a) Percent root mean square error of point estimates

Missing			Repeated MI		Bootstrapped MI	
%	Pattern	Imputations	PDMI	MLMI	PDMI	MLMI
25	MCAR	10	9.1	9.1		
		50	9.0	9.0	9.2	9.2
		200	9.0	9.0	9.0	9.0
	MAR	10	9.3	9.3		
		50	9.2	9.2	9.4	9.4
		200	9.2	9.2	9.3	9.2
50	MCAR	10	11.3	11.2		
		50	11.1	11.1	11.3	11.3
		200	11.1	11.1	11.2	11.2
	MAR	10	13.8	13.5		
		50	13.7	13.6	13.9	13.8
		200	13.5	13.5	13.5	13.5

(b) Mean length of 95% confidence intervals (CIs)
(Parentheses enclose % departure from 95% coverage)

Missing			Repeated MI					
			Score-based CIs		Within-between CIs		Bootstrapped CIs	
%	Pattern	Imputations	PDMI	MLMI	PDMI	MLMI	PDMI	MLMI
25	MCAR	10	0.18 (0.3)	0.18 (0.3)	0.18 (0.1)	0.19 (1.3)		
		50	0.18 (0.5)	0.18 (0.3)	0.18 (0.1)	0.18 (0.3)	0.19 (−0.3)	0.19 (−0.4)
		200	0.18 (0.1)	0.18 (0.1)	0.18 (0.1)	0.18 (0.1)	0.18 (−0.1)	0.18 (−0.2)
	MAR	10	0.19 (0.4)	0.18 (0.3)	0.18 (0.1)	0.19 (1.3)		
		50	0.18 (0.0)	0.18 (0.2)	0.18 (−0.1)	0.18 (0.2)	0.19 (−0.2)	0.19 (−0.3)
		200	0.18 (0.1)	0.18 (0.1)	0.18 (−0.1)	0.18 (0.0)	0.18 (−0.4)	0.18 (−0.2)
50	MCAR	10	0.23 (0.6)	0.22 (0.4)	0.23 (0.1)	0.30 (2.1)		
		50	0.22 (0.6)	0.22 (0.6)	0.22 (0.1)	0.24 (1.3)	0.24 (0.3)	0.23 (0.0)
		200	0.22 (0.4)	0.22 (0.4)	0.22 (0.0)	0.22 (0.2)	0.22 (−0.4)	0.22 (−0.4)
	MAR	10	0.28 (0.7)	0.27 (0.4)	0.29 (0.1)	0.28 (−1.7)		
		50	0.27 (0.1)	0.27 (−0.2)	0.27 (−0.2)	0.27 (−1.3)	0.29 (0.2)	0.28 (−0.4)
		200	0.27 (0.1)	0.27 (0.0)	0.26 (−0.2)	0.28 (−0.3)	0.27 (−0.4)	0.26 (−0.3)

erage. They are shorter under MLMI than under PDMI, but the difference is negligible and vanishes as the fraction of missing information gets small or the number of the imputations gets large.

WB CIs have more accurate coverage under PDMI than under MLMI. They come within 0.2% of nominal coverage under PDMI, but can drift as far as 2% above or below nominal coverage under MLMI. Coverage improves with more information or more imputations. Under most conditions, WB CIs are slightly longer, with higher coverage, under MLMI than under PDMI, but with more missing information WB CIs can be shorter under MLMI because of the shrinkage function in equation (5.18).

10.2.2 *Regression of X on Y*. We next regressed X on Y. Since the imputation model is a regression of Y on X, the imputation and analysis models are different, but both are compatible with a common bivariate normal model of (X, Y). It follows that matrix calculations should involve all 5 parameters of the bivariate normal model (see Section 10.1). To see why, let us examine what happens when matrix calculations are limited to just two parameters: the slope and intercept of the analysis model.

Table 5(a) summarizes CIs for the slope $\beta_{X.Y}$. The bootstrap and PDMI WB CIs have good coverage under all simulated conditions, but the other CIs do not. Under most simulated conditions, all CIs have good coverage, but when 50% of values are MAR, the WB intervals undercover under MLMI, and the SB intervals undercover under both PDMI and MLMI. This undercoverage does not improve as the number of imputations increases.

The reason for the undercoverage is that the SB and MLMI WB formulas have underestimated the covariance matrix of the estimates. This is because we limited the SB and PDMI WB formulas to the 2×2 covariance matrices associated with the two parameters $(\alpha_{X.Y}, \beta_{X.Y})$.

TABLE 5

Estimating the slope of X on Y. Mean length of 95% *confidence intervals* (*CIs*). (*Parentheses enclose % departure from* 95% *coverage*)

(a) With matrix formulas limited to two parameters $(\alpha_{X.Y}, \beta_{X.Y})$

Missing			Score-based CIs		Within-between CIs		Bootstrapped CIs	
%	Pattern	Imputations	PDMI	MLMI	PDMI	MLMI	PDMI	MLMI
25	MCAR	10	0.17 (0.6)	0.17 (0.5)	0.17 (0.3)	0.17 (0.9)		
		50	0.17 (0.7)	0.17 (0.6)	0.17 (0.5)	0.17 (0.3)	0.18 (0.1)	0.18 (0.5)
		200	0.17 (0.3)	0.17 (0.2)	0.17 (0.0)	0.17 (−0.2)	0.17 (−0.2)	0.17 (−0.1)
	MAR	10	0.17 (−0.2)	0.17 (−0.1)	0.17 (−0.3)	0.17 (0.1)		
		50	0.17 (0.1)	0.17 (0.0)	0.17 (−0.1)	0.17 (−0.1)	0.18 (−0.2)	0.18 (−0.4)
		200	0.17 (0.1)	0.17 (0.0)	0.17 (0.1)	0.17 (−0.4)	0.17 (−0.3)	0.17 (−0.2)
50	MCAR	10	0.20 (0.2)	0.20 (0.0)	0.20 (0.0)	0.20 (−0.2)		
		50	0.20 (0.1)	0.20 (0.0)	0.20 (0.1)	0.19 (−1.1)	0.21 (−0.2)	0.21 (−0.4)
		200	0.20 (0.5)	0.20 (0.5)	0.20 (0.4)	0.19 (−0.9)	0.20 (0.3)	0.20 (−0.1)
	MAR	10	0.21 (−1.4)	0.20 (−1.2)	0.22 (0.0)	0.20 (−1.7)		
		50	0.20 (−1.3)	0.20 (−1.5)	0.21 (0.3)	0.19 (−2.8)	0.23 (0.3)	0.23 (0.0)
		200	0.20 (−1.0)	0.20 (−1.2)	0.21 (0.5)	0.19 (−2.9)	0.21 (0.2)	0.21 (0.1)

(b) With matrix formulas including all five parameters $(\alpha_{X.Y}, \beta_{X.Y}, \sigma^2_{X.Y}, \mu_Y, \sigma^2_Y)$

Missing			Score-based CIs		Within-between CIs	
%	Pattern	Imputations	PDMI	MLMI	PDMI	MLMI
25	MCAR	10	0.17 (0.8)	0.17 (0.7)	0.17 (0.3)	0.18 (1.2)
		50	0.17 (0.8)	0.17 (0.9)	0.17 (0.5)	0.17 (0.5)
		200	0.17 (0.5)	0.17 (0.4)	0.17 (0.0)	0.17 (0.1)
	MAR	10	0.17 (0.1)	0.17 (0.2)	0.17 (−0.3)	0.18 (0.4)
		50	0.17 (0.3)	0.17 (0.3)	0.17 (−0.1)	0.17 (0.3)
		200	0.17 (0.3)	0.17 (0.3)	0.17 (0.1)	0.17 (0.0)
50	MCAR	10	0.21 (0.5)	0.20 (0.3)	0.20 (0.0)	0.21 (0.7)
		50	0.20 (0.4)	0.20 (0.2)	0.20 (0.1)	0.21 (0.4)
		200	0.20 (0.8)	0.20 (0.8)	0.20 (0.4)	0.20 (0.5)
	MAR	10	0.22 (0.8)	0.22 (0.6)	0.22 (0.0)	0.21 (−0.9)
		50	0.22 (0.6)	0.22 (0.5)	0.21 (0.3)	0.22 (0.2)
		200	0.22 (0.9)	0.21 (0.5)	0.21 (0.5)	0.22 (0.5)

But consistent estimation requires that we apply the SB and MLMI WB formulas to the full 5×5 matrix describing the five parameters of the bivariate normal model $(\mu_Y, \sigma^2_Y, \alpha_{X.Y}, \beta_{X.Y}, \sigma^2_{X.Y})$.

Table 5(b) shows what happens when we do that. The covariance matrices are now consistently estimated, and the confidence intervals have close to nominal coverage.

Although the need to use all five parameters in variance calculations is somewhat limiting, in the simulation it only made a noticeable difference when the fraction of missing information was quite large (i.e., 50% of values MAR). When the fraction of missing information was small to moderate, as it often is in applied work, neglecting parts of the parameter vector yielded acceptable results. In the next section, we will also get acceptable estimates when applying these methods to an applied dataset.

11. APPLIED DATA ANALYSIS

We next conducted an applied data analysis to compare MLMI to PDMI with different approaches to variance estimation. We analyzed data from the Millennium Cohort Study [8], a longitudinal cohort study that followed approximately 19,000 children who were born between 2000 and 2001 in the United Kingdom. We analyzed data from wave 2 of the study, when the children were around 3 years old.

Our imputation model was a general location model, which consisted of a log-linear model of the categorical variables and a conditionally multivariate normal model of the continuous variables [28]. The log-linear model included all 2-way interactions, and the mean of the multivariate normal distribution depended only on main effects of the categorical variables. The imputation model used two auxiliary variables, which were not in the analysis model but improved the imputation of variables that were [36]. One auxiliary variable was the marital status of the parents; the other was the employment status of the parent or guardian responding to the survey.

We multiply imputed missing values using both MLMI and PDMI. Under MLMI, we obtained ML parameter estimates using the EM algorithm. Under PDMI, we obtained PD parameter estimates with an MCMC algorithm that started with 100 burn-in iterations and then drew every 100th estimate from the Markov chain.

Our analysis model was a linear regression of each child's school readiness, as measured by the Bracken score, on family income, tenure of housing, any history of the child having hearing loss, ethnicity, number of siblings (categorized as 1, 2, 3+) and the age of the parent or guardian responding to the survey. The percentage of missing values varied from 0.013% for the number of siblings to 15.8% for family income. The Millennium Cohort Study uses a complex sampling scheme, but for simplicity of illustration we analyzed it as though it were a simple random sample.

For our first analysis, we used 100 imputations; for our second, we used 1000 to approximate the asymptotic behavior of the estimators. When we used repeated MI, the number of imputations was M; when we used bootstrapped MI, the number of imputations was BD, where $B = 50$ (in the first analysis) or 500 (in the second) was the number of bootstrapped samples, and $D = 2$ was the number of imputations per bootstrapped sample. We analyzed the imputed data using the linear regression model described above, applying WB, SB and bootstrap formulas to get SEs for the parameters of the analysis model.

11.1 Results with 100 Imputations

Table 6 shows results with 100 imputations. Table 6(a) gives the runtime (in seconds) needed to impute the data 100 times and analyze it on a personal computer (a 2012 MacBook Pro 2.5 GHz Intel Core i5). Although all runtimes were under a minute, imputing was much faster with MLMI than with PDMI. When we used repeated imputation, MLMI was 25 times faster than PDMI; when we used bootstrapped imputation, MLMI was 4 times faster than PDMI. Bootstrapped MLMI, though 9 times slower than repeated MLMI, was still 3 times faster than repeated PDMI. After imputation, the calculation of SEs took approximately the same runtime under MLMI as under PDMI. Score-based SE formulas were 3 times slower than other SE formulas.

The slowness of PDMI was due in part to the iterative MCMC algorithm that implemented it [28]. While MCMC is the most common PDMI algorithm, the bootstrapped EM algorithm makes PDMI faster [22], though still not as fast as MLMI.

Table 6(b) compares point estimates of the regression parameters. Among the MI estimates, the MLMI and PDMI estimates are very similar, with or without the bootstrap. This empirical result is consistent with our theoretical results showing that MI point estimates, with or without the bootstrap, are close to their asymptotic values when 100 imputations are used. The MI point estimates differ by less than 10% complete case estimates, except for the coefficient of "Other housing," which differs by a factor of 4.

Table 6(c) compares SE estimates for the regression parameters. Under repeated MI, nearly the SE estimates are very similar whether we use MLMI or PDMI, and whether we used score-based or within-between formulas.[4] This empirical result is consistent with our theoretical results showing that, with 100 imputations, score-based and within-between variance formulas come close to their asymptotic values.

[4]The one discrepancy is the SE of the "nonwhite" coefficient, which is 10% larger using the within-between formula than using the score-based formula.

TABLE 6
Imputation and analysis of Millennium Cohort Study, using 100 *imputations* ($M = 100$ *under repeated MI*, $B = 50$, $D = 2$ *under bootstrapped MI*)

(a) Runtime (in seconds)

	MLMI	PDMI	Runtime ratio (PDMI/MLMI)
Repeated imputation	1.5	37.2	25
Bootstrapped imputation	13.3	52.6	4
Within-between SE calculation	8.7	7.6	1
Score-based SE calculation	21.5	20.6	1
Bootstrap SE calculation	7.3	7.1	1

(b) Regression point estimates

	Repeated MI		Bootstrapped MI		Complete case analysis
	MLMI	PDMI	MLMI	PDMI	
Intercept	89.87	89.88	89.69	89.93	89.12
Respondent age (years)	0.24	0.24	0.24	0.24	0.27
Family income	0.89	0.89	0.90	0.89	0.87
Rented housing	−3.87	−3.89	−3.86	−3.90	−4.03
Other housing	−0.52	−0.49	−0.48	−0.42	−0.14
Child hearing loss	2.75	2.72	2.76	2.56	3.02
Nonwhite	−7.09	−7.09	−7.01	−6.99	−6.40
1 sibling	−2.42	−2.45	−2.40	−2.47	−2.52
2 siblings	−6.73	−6.74	−6.68	−6.77	−6.70
3 or more siblings	−10.78	−10.78	−10.68	−10.83	−10.59

(c) Regression standard error estimates

	Repeated MI					
	Score-based SEs		Within-between SEs		Bootstrapped SEs	
	MLMI	PDMI	MLMI	PDMI	MLMI	PDMI
Intercept	1.04	1.04	1.06	1.04	1.22	1.22
Respondent age (years)	0.02	0.02	0.02	0.02	0.02	0.03
Family income	0.04	0.04	0.04	0.04	0.05	0.05
Rented housing	0.32	0.32	0.33	0.33	0.38	0.38
Other housing	0.71	0.71	0.73	0.73	0.77	0.73
Child hearing loss	0.65	0.65	0.63	0.62	0.61	0.55
Nonwhite	0.36	0.36	0.40	0.40	0.43	0.39
1 sibling	0.31	0.31	0.31	0.31	0.36	0.26
2 siblings	0.38	0.38	0.38	0.39	0.36	0.37
3 or more siblings	0.48	0.48	0.48	0.50	0.52	0.52

Under bootstrapped MI, many of the SE estimates are similar under MLMI and PDMI, but there are a few noticeable differences. This reflects the fact that bootstrapped SE estimates can be somewhat variable when there are only $B = 50$ bootstrapped samples. With $B = 50$, the coefficient of variation for a bootstrapped SE estimates is about 10%,[5] implying that a bootstrapped SE estimate typically changes by about 10% when the data are bootstrapped and imputed again. That explains most of the differences between the bootstrapped SE estimates obtained under MLMI and PDMI. The differences do not reflect a difference between MLMI and PDMI; we would see similar differences if we had used bootstrapped MLMI twice, or bootstrapped PDMI twice. When B is larger, bootstrapped SE estimates are less variable and agree more closely under MLMI and PDMI—as we will show next.

11.2 Results with 1000 Imputations

Table 7 shows results for 1000 imputations. Table 7(a) compares runtimes. With 1000 imputations, MLMI was still much faster than PDMI. Under repeated imputation, MLMI took half of a minute, while PDMI took six and a half minutes. Under bootstrapped imputation, MLMI took two minutes, while PDMI took eight and a half. MLMI's runtime advantage of approximately six minutes was substantial, and could affect analysts' productivity and morale, especially if they re-specified the imputation model and reimputed the data several times.

Table 7(b) compares regression point estimates. The estimates are very similar under MLMI and PDMI, with or without the bootstrap. In fact, the point estimates with 1000 imputations are very close to the point estimates that we obtained with 100 imputations (Table 6(b)), confirming our claim that those point estimates were close to their asymptotic values.

Table 7(c) compares SE estimates. With 1000 imputations, nearly all the SE estimates are very similar, whether we used MLMI or PDMI with the bootstrap, the score-based formula, or the within-between formula. Evidently 1000 imputations was enough to bring all the SE estimates close to their asymptotic values. The bootstrapped SE estimates were the most variable, but with $B = 500$ they typically came within 3% of their asymptotic values.[6]

When there are substantial disagreements between different SE estimates, we favor the bootstrapped estimates because B is large and the bootstrap is consistent even when the imputation and analysis models are incompatible or misspecified. For example, for the coefficient nonwhite children, the true SE is probably closer to the 0.44-0.45 given by the bootstrap than to the 0.36 given by the SB formulas or the 0.40 given by the WB formulas. But such disagreements are rare.

[5] As discussed earlier, the coefficient of variation for an SE estimate is approximately $\sqrt{1/(2\text{df})}$, and under bootstrapped MI df is just a little smaller than B.

[6] As discussed earlier, the coefficient of variation for an SE estimate is approximately $\sqrt{1/(2\text{df})}$, and under bootstrapped MI df is just a little smaller than B. So with $B = 500$, the coefficient of variation for a bootstrapped SE estimate is 3%.

TABLE 7
Imputation and analysis of Millennium Cohort Study, using 1000 *imputations* ($M = 1000$ *under repeated MI*, $B = 500$, $D = 2$ *under bootstrapped MI*)

(a) Runtime (in seconds)

	MLMI	PDMI	Runtime ratio (PDMI/MLMI)
Repeated imputation	33.4	394.5	12
Bootstrapped imputation	123.6	512.4	4
Within-between SE calculation	78.2	75.2	1
Score-based SE calculation	217.0	214.5	1
Bootstrap SE calculation	64.5	65.9	1

(b) Regression point estimates

	Repeated MI		Bootstrapped MI		Complete case analysis
	MLMI	PDMI	MLMI	PDMI	
Intercept	89.85	89.84	89.84	89.81	89.12
Respondent age (years)	0.24	0.24	0.24	0.24	0.27
Family income	0.89	0.89	0.89	0.90	0.87
Rented housing	−3.86	−3.86	−3.86	−3.87	−4.03
Other housing	−0.50	−0.48	−0.53	−0.50	−0.14
Child hearing loss	2.73	2.72	2.73	2.69	3.02
Nonwhite	−7.08	−7.06	−7.08	−7.06	−6.40
1 sibling	−2.44	−2.45	−2.44	−2.43	−2.52
2 siblings	−6.74	−6.74	−6.74	−6.76	−6.70
3 or more siblings	−10.78	−10.77	−10.72	−10.77	−10.59

(c) Regression standard error estimates

	Score-based SEs		Within-between SEs		Bootstrapped SEs	
	MLMI	PDMI	MLMI	PDMI	MLMI	PDMI
Intercept	1.04	1.04	1.03	1.03	1.03	1.04
Respondent age (years)	0.02	0.02	0.02	0.02	0.03	0.02
Family income	0.04	0.04	0.04	0.04	0.05	0.05
Rented housing	0.32	0.32	0.33	0.33	0.35	0.37
Other housing	0.71	0.71	0.72	0.71	0.73	0.71
Child hearing loss	0.65	0.65	0.62	0.62	0.60	0.59
Nonwhite	0.36	0.36	0.40	0.40	0.44	0.45
1 sibling	0.31	0.31	0.31	0.31	0.32	0.34
2 siblings	0.38	0.38	0.39	0.39	0.41	0.40
3 or more siblings	0.48	0.48	0.48	0.48	0.51	0.51

How surprised should we be that the different SE formulas agree so well? There are two considerations. First, the formulas make different assumptions about the imputation and analysis models (Section 7).

- The bootstrap SE formulas are consistent even when the imputation and analysis models are incompatible or misspecified. So they are consistent here.
- The PDMI WB formula is consistent when the imputation and analysis models are compatible and correct. Here, the imputation and analysis models are compatible [7], and although they are unlikely to be perfectly specified, evidently any misspecification is not serious enough to introduce much bias. If there were much bias, we would more often see the PDMI WB SEs disagreeing with the bootstrap.
- The SB and MLMI WB matrix formulas have additional requirements. Not only must the imputation and analysis models be consistent and correct, but the SB and MLMI WB matrices should include parameters from the imputation model that are not in the analysis model. In this example, though, the matrices included only parameters from the analysis model—and returned SE estimates that were mostly similar to the consistent bootstrapped estimates.

Perhaps a reason for the near-agreement across different formulas is that the fraction of missing information is rather small. In our simulations, we found that the differences among SE estimates were barely noticeable unless the fraction of missing information was quite large.

12. CONCLUSION

MLMI offers a serious alternative to PDMI. MLMI is not the only alternative—fractional imputation also deserves serious consideration [38]—but it does have certain advantages over PDMI.

The first advantage of MLMI is its computational efficiency. MLMI is easier to code than PDMI, and MLMI runs faster: it can produce more imputations in the same runtime. The speed advantage of MLMI is substantial when PDMI uses MCMC to get posterior draws, as most PDMI software does. The speed advantage of MLMI is more modest when PDMI gets posterior draws with a more efficient algorithm, such as bootstrapped ML [22].

The second advantage of MLMI is the efficiency of its point estimates. Compared to PDMI point estimates, MLMI point estimates are more efficient when they use the same number of imputations as PDMI, and still more efficient when MLMI uses the larger number of imputations that it can generate in the same runtime as PDMI. The efficiency advantage of MLMI point estimates is typically quite small, but can be larger when the fraction of missing information is large and PDMI uses few imputations.

Until now, the use of MLMI has been discouraged by the lack of convenient formulas for variances, SEs, and CIs. But we have derived and evaluated three SE estimators: the within-between (WB) estimator, the score-based (SB) estimator and bootstrapped MI. Some of these SE estimators are more viable than others.

The WB variance formulas use variance components that lie within and between the imputed datasets. An old

WB formula (5.5) has been used with PDMI for over 30 years [27], and we have derived a new WB formula (5.16) that is consistent under MLMI. Our MLMI WB formula requires more imputations than the PDMI WB imputations, but when the fraction of missing information is 50% or less, the number of imputations required is not excessive and often present no practical problem since MLMI produces imputations more quickly than PDMI (Table 3). With more than 50% missing information, though, the MLMI WB formula requires a rapidly increasing number of imputations, so that it becomes better to use bootstrapped MI, which with high missing information can produce better SE estimates with fewer imputations.

The SB variance formulas decompose the variance of the score function. The same SB formula is consistent under PDMI and MLMI. The SB variance formula needs fewer imputations than the WB formulas, but its calculation requires the contribution of each case to the score function. This can be a serious disadvantage, since the user often does not know the contribution of each case to the score function, and some approaches to estimation do not use the score function at all. This limits the practical use of the SB formula.

When the imputation and analysis models are the same, both the SB formula and the MLMI WB formula can be applied to the parameters of the analysis model alone. But when the imputation and analysis models are different, the SB and MLMI WB formulas can also require the parameters of the underlying common model that is consistent with both the imputation and analysis model. When these additional parameters are neglected, the SB and MLMI WB formulas can produce poor SE estimates, although in practice the SE estimates seem to perform well unless the fraction of information is quite large.

Bootstrapped MI variance estimation is the most robust approach. It is flexible and can work with a variety of imputation methods, including but not limited to PDMI and MLMI. Bootstrapped MI variance estimates are consistent even when the imputation and analysis models are different or misspecified. Unlike SB and MLMI WB estimates, bootstrapped MI estimates never require parameter estimates beyond those from the analysis model. Unlike WB variance estimates, bootstrapped MI variance estimates do not require a complete-data analytic SE for complete data, and so can be used in situations where analytic SEs are unavailable or invalid.

A further advantage of bootstrapped MI variance estimates is that they are consistent even when the imputation and analysis models are incompatible or misspecified. This property is valuable since in practical settings most models are at least a little misspecified, and incompatibility between the imputation and analysis models is common. While no method can ensure that *point estimates* will be consistent under a misspecified model, bootstrapped MI can at least ensure that the variability of point estimates is estimated accurately. This is a property that WB and SB estimates lack, under both MLMI and PDMI. We know of only one other approach that can produce consistent variance estimates under misspecified and incompatible imputation and analysis models [24]—but the calculations are relatively complicated and require statistics, including but not limited to the score function, that users often lack access to in practical settings.

Bootstrapped MI, by contrast, is straightforward. An old knock against bootstrapped MI was that it seemed to require a large number of imputations D for each bootstrapped sample [10]. Our approach, however, produces consistent variance estimates with just $D = 2$ imputations. Another knock was that the bootstrap can require a large number of bootstrapped samples B, but that requirement is not limited to imputed data. In complete data, the bootstrap can also require a large B, and analysts often consider that an acceptable price to pay for robust SE estimates. Bootstrapping MI requires approximately the same B as bootstrapping complete data. In both complete and MI data, the degrees of freedom is slightly less than B, and perhaps $B = 25$ samples suffice for replicable point estimates, and $B = 500$ for replicable SE estimates. Imputing B bootstrapped samples can take a long time if you use PDMI, but MLMI can impute the bootstrapped samples much more quickly.

APPENDIX A: SIMPLIFIED EXPRESSION FOR V_{PDMI}

In equation (4.5) we gave an expression for V_{PDMI} which we claimed was equivalent to the more complicated expression in equation (2) from Wang and Robins [37]. Below we give the steps of the simplification. The first line gives equation (2) from Wang and Robins [37], with a typo corrected and the symbols changed to match our notation. The last line gives our simplified expression (4.5).

$$
\begin{aligned}
V_{\text{PDMI}} &= V_{\text{ML}} + \frac{1}{M} V_{\text{com}} \gamma_{\text{mis}} + \frac{1}{M} \gamma_{\text{mis}}^T V_{\text{ML}} \gamma_{\text{mis}} \\
&= V_{\text{ML}} + \frac{1}{M} (V_{\text{com}} + \gamma_{\text{mis}}^T V_{\text{ML}}) \gamma_{\text{mis}} \\
&= V_{\text{ML}} + \frac{1}{M} (V_{\text{com}} + (I - \gamma_{\text{obs}})^T V_{\text{ML}}) \gamma_{\text{mis}} \\
&= V_{\text{ML}} + \frac{1}{M} (V_{\text{com}} + (I - V_{\text{ML}}^{-1} V_{\text{com}})^T V_{\text{ML}}) \gamma_{\text{mis}} \\
&= V_{\text{ML}} + \frac{1}{M} (V_{\text{com}} \\
&\quad + ((V_{\text{ML}}^{-1} (V_{\text{ML}} - V_{\text{com}}))^T V_{\text{ML}}) \gamma_{\text{mis}} \\
&= V_{\text{ML}} + \frac{1}{M} (V_{\text{com}} \\
&\quad + (V_{\text{ML}} - V_{\text{com}})^T V_{\text{ML}}^{-T} V_{\text{ML}}) \gamma_{\text{mis}}
\end{aligned}
$$

$$= V_{\mathrm{ML}} + \frac{1}{M}(V_{\mathrm{com}} + (V_{\mathrm{ML}} - V_{\mathrm{com}})V_{\mathrm{ML}}^{-1}V_{\mathrm{ML}})\gamma_{\mathrm{mis}}$$

$$= V_{\mathrm{ML}} + \frac{1}{M}(V_{\mathrm{com}} + V_{\mathrm{ML}} - V_{\mathrm{com}})\gamma_{\mathrm{mis}}$$

$$= V_{\mathrm{ML}} + \frac{1}{M}V_{\mathrm{ML}}\gamma_{\mathrm{mis}}.$$

APPENDIX B: SHRINKING WB ESTIMATES UNDER MLMI

In Section 5.2 we presented a simple estimator $\widehat{\gamma}_{\mathrm{mis|MLMI,WB}} = \widehat{W}_{\mathrm{MLMI}}^{-1}\widehat{B}_{\mathrm{MLMI}}$ for the fraction of missing information under MLMI, then replaced it with the shrunken estimator $\widetilde{\gamma}_{\mathrm{mis|MLMI}} = h(\widehat{\gamma}, M-1)$. We now explain why shrinkage is necessary, and justify our shrinkage function $h()$.

The problem with the simple estimator $\widehat{\gamma}_{\mathrm{mis|MLMI,WB}}$ is that it can exceed 1, whereas the true fraction of missing information γ_{mis} cannot. To show this, we adopt the convention, common in the MI literature, that the variation in $\widehat{W}_{\mathrm{MI}}$ is negligible compared to the variation in $\widehat{B}_{\mathrm{MI}}$. Then the distribution of $\widehat{\gamma}_{\mathrm{mis|MLMI}}$ is approximately scaled chi-square:

$$\widehat{\gamma}_{\mathrm{mis|MLMI}} = \gamma_{\mathrm{mis}}\frac{U}{M-1} \quad \text{where } U \sim \chi^2_{M-1} \tag{B.1}$$

and the probability that $\widehat{\gamma}_{\mathrm{mis|MLMI}}$ exceeds 1 is $P(\gamma_{\mathrm{mis}}\frac{U}{M-1} > 1) = P(U > \frac{M-1}{\gamma_{\mathrm{mis}}})$. Figure 1 graphs this probability as a function of M and γ_{mis}. The probability is negligible if γ_{mis} is low, but can be substantial if γ_{mis} is high and M is low relative to γ_{mis}.

Our solution is to replace $\widehat{\gamma}_{\mathrm{mis|MLMI,WB}}$ with a shrunken estimator $\widetilde{\gamma}_{\mathrm{mis|MLMI,WB}}$ which is guaranteed to take values in $(0, 1)$. We define $\widetilde{\gamma}_{\mathrm{mis|MLMI,WB}}$ as the posterior mean of γ_{mis} when the prior is uniform on $(0, 1)$. With this prior, the posterior distribution of γ_{mis} approximates a scaled inverse chi-square—

$$\gamma_{\mathrm{mis}} = \widehat{\gamma}_{\mathrm{mis,|MLMI,WB}}\frac{M-1}{U} \quad \text{where } U \sim \chi^2_{M-1} \tag{B.2}$$

—with the modification that the distribution of γ_{mis} is truncated on the right at 1. We calculated the mean of this truncated distribution using Mathematica software, version 8. The solution is (5.18), that is,

$$\widetilde{\gamma}_{\mathrm{mis|MLMI,WB}} = h(\widehat{\gamma}_{\mathrm{mis|MLMI,WB}}, M-1), \tag{B.3}$$

where

$$h(\widehat{\gamma}, \nu) = \frac{\nu}{2}\widehat{\gamma}\frac{\Gamma(\frac{\nu-2}{2}, \frac{\nu}{2}\widehat{\gamma})}{\Gamma(\frac{\nu}{2}, \frac{\nu}{2}\widehat{\gamma})}. \tag{B.4}$$

Using numerical integration in Mathematica software, we calculate the bias $E(\widetilde{\gamma}_{\mathrm{mis|MLMI,WB}} - \gamma_{\mathrm{mis}})$ that is summarized in Table 2.

Since the function $\Gamma(a, z)$ is unavailable in some statistical software, for implementation purposes it helps to know that with $\nu > 2$, $h(\widehat{\gamma}, \nu)$ simplifies to

$$h(\widehat{\gamma}, \nu) = \frac{\nu}{\nu-2}\widehat{\gamma}\frac{R_\Gamma(\frac{\nu-2}{2}, \frac{\nu}{2}\widehat{\gamma})}{R_\Gamma(\frac{\nu}{2}, \frac{\nu}{2}\widehat{\gamma})}, \tag{B.5}$$

where $R_\Gamma(a, z)$, which is widely available in statistical software, is the survival function for a gamma distribution with shape parameter a, evaluated at z. Since this simplification requires $\nu > 2$, it can only be used when $M > 4$.

APPENDIX C: DEGREES OF FREEDOM FOR WB VARIANCE ESTIMATION UNDER MLMI

Equation (5.21) approximates the df of the variance estimate $\widetilde{V}_{\mathrm{MLMI,WB}}$. Although $\widetilde{V}_{\mathrm{MLMI,WB}}$ is not a chi-square variable, a chi-squared variable with $\mathrm{df} = \widehat{\nu}_{\mathrm{MLMI,WB}}$ will have approximately the same coefficient of variation (CV) as $\widetilde{V}_{\mathrm{MLMI,WB}}$.

To derive this approximation, consider the scalar expression

$$\widetilde{V}_{\mathrm{MLMI,WB}} = \widetilde{V}_{\mathrm{ML|MLMI,WB}} + \frac{1}{M}\widehat{B}_{\mathrm{MLMI}}, \tag{C.1}$$

where

$$\widetilde{V}_{\mathrm{ML|MLMI,WB}} = \widehat{W}_{\mathrm{MLMI}}\widetilde{\gamma}^{-1}_{\mathrm{obs|MLMI,WB}}, \tag{C.2}$$

$$\widetilde{\gamma}_{\mathrm{obs|MLMI,WB}} = 1 - \widetilde{\gamma}_{\mathrm{mis|MLMI,WB}}, \tag{C.3}$$

$$\widetilde{\gamma}_{\mathrm{mis|MLMI,WB}} = h(\widehat{\gamma}_{\mathrm{mis|MLMI,WB}}), \tag{C.4}$$

$$\widehat{\gamma}_{\mathrm{mis|MLMI,WB}} = \widehat{W}^{-1}_{\mathrm{MLMI}}\widehat{B}_{\mathrm{MLMI}}. \tag{C.5}$$

We can approximate the distribution of $\widetilde{V}_{\mathrm{MLMI,WB}}$ by starting with its components. $\widehat{B}_{\mathrm{MLMI}}$ has approximately a scaled χ^2_{M-1} distribution, and if we regard $\widehat{W}_{\mathrm{MLMI}}$ as fixed, then $\widehat{\gamma}_{\mathrm{mis|MLMI,WB}}$ also has approximately a scaled χ^2_{M-1} distribution with expectation γ_{mis}. We regard $\widetilde{\gamma}_{\mathrm{mis|MLMI,WB}}$ as having approximately the same distribution as $\widehat{\gamma}_{\mathrm{mis|MLMI,WB}}$.

Under these assumptions, $\widetilde{\gamma}_{\mathrm{obs|MLMI,WB}}$ has expectation γ_{obs}, standard deviation $\gamma_{\mathrm{mis}}\sqrt{2/(M-1)}$, and $\mathrm{CV} = (\frac{\gamma_{\mathrm{mis}}}{\gamma_{\mathrm{obs}}})\sqrt{2/(M-1)}$, which is also the CV of a $\chi^2_{\nu_1}$ variable with $\mathrm{df} = \nu_1 = (M-1)(\frac{\gamma_{\mathrm{obs}}}{\gamma_{\mathrm{mis}}})^2$. So we can approximate $\widetilde{\gamma}_{\mathrm{obs|MLMI,WB}}$ as a scaled $\chi^2_{\nu_1}$ variable.

Then $\widetilde{\gamma}^{-1}_{\mathrm{obs|MLMI,WB}}$ approximates a scaled inverse chi-square variable with $\mathrm{df} = \nu_1$, but this inverse chi-square has the same CV as an ordinary chi-square variable with $\mathrm{df} = \nu_1 - 4$. So we can approximate $\widetilde{\gamma}^{-1}_{\mathrm{obs|MLMI,WB}}$ as a scaled $\chi^2_{\nu_1-4}$ variable. It follows that $\widetilde{V}_{\mathrm{MLMI,WB}}$ is approximately scaled $\chi^2_{\nu_1-4}$ as well.

Now

$$\widetilde{V}_{\mathrm{MLMI,WB}} = \widetilde{V}_{\mathrm{ML|MLMI,WB}} + \frac{1}{M}\widehat{B}_{\mathrm{MLMI}} \tag{C.6}$$

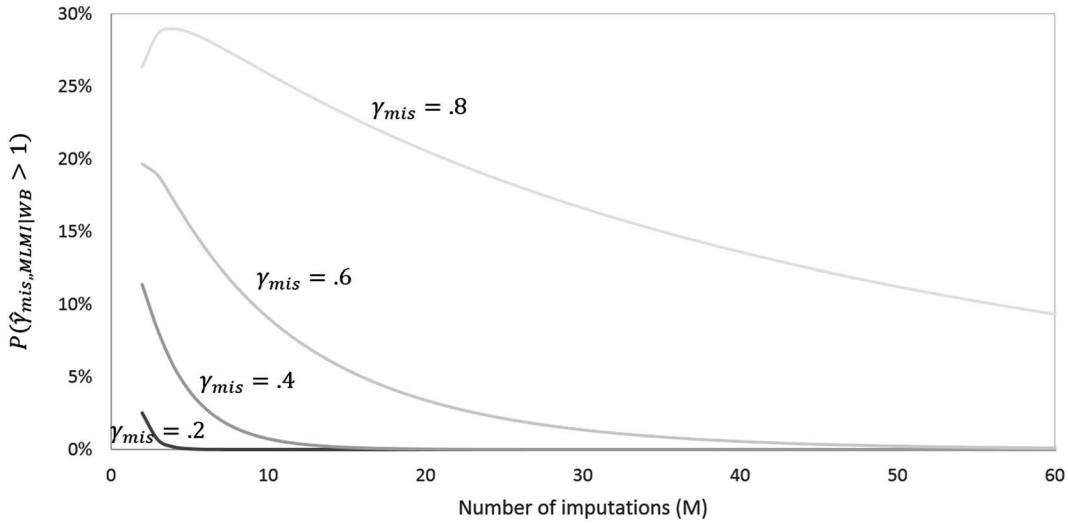

FIG. 1. *The probability that* $\hat{\gamma}_{\text{mis|MLMI}}$ *exceeds* 1, *as a function of* M *and* γ_{mis}.

is the sum of two scaled chi-square variables with respective dfs equal to $\nu_1 - 4$ and $M - 1$. The variables are not independent, but the covariance between them is negligible if M is large or γ_{mis} is small. If we apply the Satterthwaite approximation to the sum, we get expression (5.21) for the df of $\widetilde{V}_{\text{MLMI,WB}}$.

APPENDIX D: WANG & ROBINS' SB ESTIMATORS

In Section 6 we mentioned that Wang and Robins [37], Lemma 2, use a different SB estimator for V_{ML}^{-1}. After correction of a typo,[7] their estimator is

$$\text{(D.1)} \qquad \check{V}_{\text{ML|SB}}^{-1} = \frac{1}{M(M-1)} \sum_{m \neq m'} \sum_{i=1}^{N} c_{mm',i},$$

where

$$\text{(D.2)} \quad c_{mm',i} = \frac{1}{2}\left(\hat{s}_{\text{com},m,i}^{T} \hat{s}_{\text{com},m',i} + \hat{s}_{\text{com},m',i}^{T} \hat{s}_{\text{com},m,i}\right)$$

is the "symmetrized" cross-product of score estimates between one SI dataset (m) and another (m'). The cross-product $\hat{s}_{\text{com},m,i}^{T} \hat{s}_{\text{com},m',i}$ is not symmetric, and neither is the reverse cross-product $\hat{s}_{\text{com},m',i}^{T} \hat{s}_{\text{com},m,i}$, but the average $c_{mm'}$ is symmetric and so can be used to estimate the symmetric matrix V_{ML}^{-1}.

Since $c_{mm'} = c_{m'm}$ we can halve the number of cross-products we need to calculate by restricting ourselves to cross-products where $m < m'$. Then Wang and Robins' estimator simplifies to

$$\text{(D.3)} \qquad \check{V}_{\text{ML|SB}}^{-1} = \frac{2}{M(M-1)} \sum_{m < m'} \sum_{i=1}^{N} c_{mm',i}.$$

[7] Wang and Robins inadvertently divide $V_{\text{ML|SB}}^{-1}$ by N.

$\check{V}_{\text{ML|SB}}^{-1}$ looks quite different from our estimator $\hat{V}_{\text{ML|SB}}^{-1}$, but in fact the two are just different formulas for estimating the between-group variance of $\hat{s}_{\text{com},m,i}$. To see this, notice that, if $\hat{s}_{\text{com},m,i}$ is scalar, then $\check{V}_{\text{ML|SB}}^{-1}$ becomes

$$\text{(D.4)} \quad \check{V}_{\text{ML|SB}}^{-1} = \frac{2}{M(M-1)} \sum_{m < m'} \sum_{i=1}^{N} \hat{s}_{\text{com},m,i} \hat{s}_{\text{com},m',i}$$

which, if divided by N and $V(\hat{s}_{\text{com},m,i})$, is just a century-old formula for estimating the intraclass correlation [13, 14].[8] The intraclass correlation formula can be simplified so that no cross-products are required [14]; applying the simplification, we get

$$\text{(D.5)} \quad \check{V}_{\text{ML|SB}}^{-1} = \frac{M}{M-1} \sum_{m < m'} \sum_{i=1}^{N} (\overline{s}_{\text{com},i})^{\otimes 2} - \frac{1}{M-1} \hat{V}_{\text{com|SB}}^{-1}$$

which is very similar to our $\hat{V}_{\text{ML|SB}}^{-1}$.

REFERENCES

[1] ANDERSON, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Assoc.* **52** 200–203. MR0087286

[2] ARBUCKLE, J. L. (1996). Full information estimation in the presence of incomplete data. In *Advanced Structural Equation Modeling*: *Issues and Techniques* 243–277.

[3] BARNARD, J. and RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86** 948–955. MR1741991 https://doi.org/10.1093/biomet/86.4.948

[8] The old formula would center $\hat{s}_{\text{com},m,i}$ around its sample mean, but that is not necessary here since we know that the mean of $\hat{s}_{\text{com},m,i}$ is zero.

[4] BARTLETT, J. W. (2019). bootImpute: Bootstrap inference for multiple imputation. Comprehensive R Archive Network. https://CRAN.R-project.org/package=bootImpute.

[5] BARTLETT, J. W. (2019). mlmi: Maximum likelihood multiple imputation. Comprehensive R Archive Network. https://CRAN.R-project.org/package=mlmi.

[6] BARTLETT, J. W. and KEOGH, R. H. (2019). smcfcs: Multiple imputation of covariates by substantive model compatible fully conditional specification. Comprehensive R Archive Network. https://CRAN.R-project.org/package=smcfcs.

[7] BARTLETT, J. W., SEAMAN, S. R., WHITE, I. R. and CARPENTER, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat. Methods Med. Res.* **24** 462–487. MR3372102 https://doi.org/10.1177/0962280214521348

[8] CENTRE FOR LONGITUDINAL STUDIES, INSTITUTE OF EDUCATION, UNIVERSITY OF LONDON (2017). Millennium cohort study: Second survey 2003-2005, UK Data Service, 9th edition. SN: 5350. https://doi.org/10.5255/UKDA-SN-5350-4

[9] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

[10] EFRON, B. (1994). Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.* **89** 463–479. MR1294072

[11] EID, S. (2016). Mult[i]ple Imputation taking forever!! Retrieved May 8, 2017, from http://www.statalist.org/forums/forum/general-stata-discussion/general/1330365-multple-imputation-taking-forever.

[12] ENDERS, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Struct. Equ. Model.* **8** 128–141.

[13] FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, London. MR0346954

[14] HARRIS, J. A. (1913). On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika* **9** 446–472. https://doi.org/10.2307/2331901

[15] HEITJAN, D. F. and LITTLE, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **40** 13–29.

[16] HONAKER, J. and KING, G. (2010). What to do about missing values in time-series cross-section data. *Amer. J. Polit. Sci.* **54** 561–581.

[17] HONAKER, J., KING, G. and BLACKWELL, M. (2015). AMELIA II: A Program for missing data, version 1.7.4.

[18] HUANG, J. (2015). How to speed up multiple imputation process. Retrieved May 8, 2017, from http://www.statalist.org/forums/forum/general-stata-discussion/general/1305705-how-to-speed-up-multiple-imputation-process.

[19] SAS INSTITUTE (2000). The MI procedure for SAS version 8.1. Cary, NC.

[20] KIM, J. K. (2004). Finite sample properties of multiple imputation estimators. *Ann. Statist.* **32** 766–783. MR2060177 https://doi.org/10.1214/009053604000000175

[21] KIM, J. K. and RAO, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika* **96** 917–932. MR2767279 https://doi.org/10.1093/biomet/asp041

[22] KING, G., HONAKER, J., JOSEPH, A. and SCHEVE, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am. Polit. Sci. Rev.* **95** 49–69.

[23] LANNING, D. and BERRY, D. (2003). An alternative to PROC MI for large samples (SUGI 28-271). In 28*th Meeting of the SAS Users Group International*, *Seattle*, *WA*. Retrieved from http://www2.sas.com/proceedings/sugi28/271-28.pdf.

[24] ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. MR1766832 https://doi.org/10.1093/biomet/87.1.113

[25] ROJAS, F. (2012). mi impute: a stata command review. Retrieved May 8, 2017, from https://orgtheory.wordpress.com/2012/02/17/mi-impute-a-stata-command-review/.

[26] RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

[27] RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

[28] SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. *Monographs on Statistics and Applied Probability* **72**. CRC Press, London. MR1692799 https://doi.org/10.1201/9781439821862

[29] SOCIAL SCIENCE COMPUTING COOPERATIVE, UNIVERSITY OF WISCONSIN (2012). Speeding up Multiple Imputation in Stata using Parallel Processing. Retrieved May 8, 2017, from https://www.ssc.wisc.edu/sscc/pubs/stata_mi_condor.htm.

[30] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

[31] VAN BUUREN, S. (2018). *Flexible Imputation of Missing Data*, 2nd ed. CRC Press/CRC, Boca Raton, FL.

[32] VAN BUUREN, S., GROOTHUIS-OUDSHOORN, K., ROBITZSCH, A., VINK, G., DOOVE, L., JOLANI, S., SCHOUTEN, R., GAFFERT, P., MEINFELDER, F. et al. (2018). Mice: Multivariate imputation by chained equations. Comprehensive R Archive Network. https://CRAN.R-project.org/package=mice.

[33] VON HIPPEL, P. T. (2013). The bias and efficiency of incomplete-data estimators in small univariate normal samples. *Sociol. Methods Res.* **42** 531–558. MR3190739 https://doi.org/10.1177/0049124113494582

[34] VON HIPPEL, P. T. (2016). New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples. *Struct. Equ. Model.* **23** 422–437. MR3488832 https://doi.org/10.1080/10705511.2015.1047931

[35] VON HIPPEL, P. T. (2018). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociol. Methods Res.* https://doi.org/10.1177/0049124117747303

[36] VON HIPPEL, P. T. and LYNCH, J. (2013). Efficiency gains from using auxiliary variables in imputation. arXiv preprint, arXiv:1311.5249.

[37] WANG, N. and ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948. MR1666715 https://doi.org/10.1093/biomet/85.4.935

[38] YANG, S. and KIM, J. K. (2016). Fractional imputation in survey sampling: A comparative review. *Statist. Sci.* **31** 415–432. MR3552742 https://doi.org/10.1214/16-STS569