The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Iterative Multiple Imputation: A Framework to Determine the Number of Imputed Datasets

Vahid Nassiri, Geert Molenberghs, Geert Verbeke & João Barbosa-Breda

To cite this article: Vahid Nassiri, Geert Molenberghs, Geert Verbeke & João Barbosa-Breda (2018): Iterative Multiple Imputation: A Framework to Determine the Number of Imputed Datasets, The American Statistician, DOI: 10.1080/00031305.2018.1543615

To link to this article: https://doi.org/10.1080/00031305.2018.1543615

THE AMERICAN STATISTICIAN
2019, VOL. 00, NO. 0, 1–10: Statistical Practice
https://doi.org/10.1080/00031305.2018.1543615

Iterative Multiple Imputation: A Framework to Determine the Number of Imputed Datasets

Vahid Nassiri[a], Geert Molenberghs[a,b], Geert Verbeke[a,b], and João Barbosa-Breda[c,d,e]

[a]I-BioStat, KU Leuven, Leuven, Belgium; [b]I-BioStat, Universiteit Hasselt, Hasselt, Belgium; [c]Ophthalmology Department, Centro Hospitalar São João, Porto, Portugal; [d]Surgery and Physiology Unit, Faculty of Medicine of the University of Porto, Porto, Portugal; [e]Research Group Ophthalmology, KU Leuven, Leuven, Belgium

ABSTRACT

We consider multiple imputation as a procedure iterating over a set of imputed datasets. Based on an appropriate stopping rule the number of imputed datasets is determined. Simulations and real-data analyses indicate that the sufficient number of imputed datasets may in some cases be substantially larger than the very small numbers that are usually recommended. For an easier use in various applications, the proposed method is implemented in the R package `imi`.

ARTICLE HISTORY
Received November 2016
Accepted October 2018

KEYWORDS
Central limit theorem; Incomplete data; Iterative procedure; Missing data.

1. Introduction

Since Rubin's seminal work on multiple imputation (MI, Rubin 1978, 1979, 1987), the method has been broadly applied, methodologically extended, and expanded toward ever more areas of application (Carpenter and Kenward 2008; van Buuren 2012; Carpenter and Kenward 2012; Van der Elst et al. 2015). MI is a commonly used approach to analyze incomplete data. A growing literature and increasing number of software implementations have contributed to the spread of the method. One of the attractions of MI is its very good to excellent performance even with a relatively small number of imputed datasets. This was important when Rubin created the method, about 40 years ago, in view of, among others, the US Census. It still is today because of ever increasing data streams.

Broadly, MI replaces a missing value with several plausible values, sampled from an appropriate predictive distribution for the missing values, given observed information. The method produces several completed datasets to replace the initial partially observed dataset.

An evident practical question is how many imputed datasets, M say, are sufficient for reliable results, knowing that full efficiency would be reached for $M = +\infty$. Precision should be balanced against computational expense. It has been stated repeatedly, and practically confirmed, that a small number of imputations oftentimes gives very acceptable results. Sources like Rubin (1987) and the classic Rubin and Little (2002) quote values as low as 2–5. While attractive, especially when faced with large and complex databases, such low numbers may not always apply. Features that would require larger numbers of imputation include: increasing amounts of missing information, and the need for hypothesis testing rather than merely parameter and precision estimation. It is reassuring that Schafer (1997) indicated that, even under undesirable circumstances, it is rare for M needing to be in excess of 20. This observation was based on extensive simulation, rather than theory.

Some of the approaches for dealing with selecting the number of imputations are reviewed in Section 2. Our proposal to handle the choice is presented in Section 3. Section 4 reports some simulations, while real-life applications are offered in Section 5.

2. Number of Imputed Datasets: A Review

Upon producing multiply imputed datasets, a standard analysis is routinely applied to each of the completed datasets. Let $\boldsymbol{\theta}$ be the parameter vector of interest and $\widehat{\boldsymbol{\theta}}_m$ its estimate from the mth imputed dataset. If M is the number of imputed datasets, then $\tilde{\boldsymbol{\theta}}$, the MI estimator is defined as

$$\widetilde{\boldsymbol{\theta}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\theta}}_m. \tag{1}$$

If $\widehat{\Sigma}_m$ is the estimated variance-covariance of $\widehat{\boldsymbol{\theta}}_m$, then the within-imputation variability is

$$\widehat{W} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\Sigma}_m, \tag{2}$$

and the between-imputation variability is

$$\widehat{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}\right)'. \tag{3}$$

Then, asymptotically, for the sample size as well as M going to infinity, we have

CONTACT Vahid Nassiri *vahid.nassiri@kuleuven.be* I-BioStat, KU Leuven, B-3000 Leuven, Belgium.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.

$$\widetilde{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \widehat{B} + \widehat{W}). \tag{4}$$

For finite M, the variance of the normal distribution in (4) becomes (Rubin and Little 2002)

$$\widehat{V} = (1 + M^{-1})\widehat{B} + \widehat{W}. \tag{5}$$

The term $\widehat{B}/M$ takes into account the increased variability stemming from finite M. Obviously, for $M \to \infty$, this extra term vanishes. Consider $r = \frac{1}{q}\big(\text{tr}(\widehat{B}\widehat{W}^{-1})\big)(1 + M^{-1})$ and $\nu = (M-1)(1+r^{-1})^2$ with $\text{tr}(A)$ indicating the trace of the matrix A, and q the length of parameter vector $\boldsymbol{\theta}$. For the kth element of $\boldsymbol{\theta}$ we have

$$\widetilde{\theta}_k \approx \theta_k + t_\nu \sqrt{\widehat{V}_{kk}}, \tag{6}$$

where t_ν is Student's t distribution with ν degrees-of-freedom and $\widehat{V}$ can be computed using (5). These expressions are derived and discussed in detail by Rubin (1987) or Carpenter and Kenward (2012). The degrees-of-freedom ν were computed by Rubin and Schenker (1986). Small sample degrees-of-freedom were reviewed by Wagstaff and Harel (2011). Note that (6) applies to the scalar case.

Now, if I_M is the information based on M imputations and I_∞ is the information for $M \to +\infty$, then,

$$\frac{I_M}{I_\infty} = \left(1 + \frac{\gamma}{M}\right)^{-1}, \quad \gamma = \frac{r + 2/(\nu+3)}{r+1}, \tag{7}$$

where γ in (7) can be regarded as the fraction of missing information. Rubin (1987) suggested that for many applications with a moderate amount of missingness, 3–5 imputations might well be sufficient. For example, using these expressions, with 10% missingness, $M = 3$ would provide about 97% efficiency when compared with an infinite number of imputed datasets. It is not surprising that this rule of thumb is relatively broadly accepted in practice. For example, according to SAS/STAT(R) 9.4 User's Guide, the default number of imputations in the MI procedure is set equal to 5.

However, in the latest version of SAS (SAS/STAT(R) 14.1) the default number of imputed datasets (`NIMPUTE`) is increased to 25. Furthermore, following White, Royston, and Wood (2011), an option is provided in this version of SAS' `PROC MI` to set the number of imputed datasets equal to the fraction of incomplete cases (`NIMPUTE=PCTMISSING`). This indicates a level of acceptance among commercial software developers for the need for larger M's, also alternative procedures to determine it.

A practical difficulty with this heuristic is the need for γ. The quantity should be estimated to begin with, but obviously, the quality of this estimate is hugely affected by the number of imputed datasets itself. Furthermore, as Bodner (2008) also suggested, the rule for the number of imputed datasets would take into account three parameters: $\widetilde{\boldsymbol{\theta}}$, its variance, and the degrees-of-freedom of the Student's t distribution. When one is interested in controlling the p-values when conducting hypothesis tests, or other quantities, then perhaps more imputations are needed (Harel and Schafer 2003; Carpenter and Kenward 2008).

Carpenter and Kenward (2012) distinguished between two cases. Their proposal is to use a small number of imputed datasets when the inference is clear-cut, but if it is less clear-cut and an accurate estimate of the p-value or γ is needed, they proposed to set $M = 100$.

Given that for small M, the approximation in (6) may not be accurate enough, Royston (2004) proposed an iterative procedure to select M such that the confidence level remains at a selected level. He proposed to select M such that the coefficient of variation of $t_\nu\sqrt{\widehat{V}_{kk}}$ for the worst-case parameter becomes less than the Type I error rate α; this author used the conventional $\alpha = 0.05$.

Graham, Olchowski, and Gilreath (2007) performed a simulation study to investigate the effect of M on various characteristics, such as power, mean squared error, and fraction of missingness. Bodner (2008) explored various factors that are affected by increasing M. His findings were used in Lu (2017) who considered the case of longitudinal data and then focused on determining M to stabilize MI-based inferences. Stability in these articles is defined in terms of four conditions regarding the conditional SEs of the MI estimator, the test statistics, the missingness fraction, as well as the coefficient of variation of the half confidence interval. Under some circumstances, these authors proposed to use as many as 200 imputed datasets.

Royston, Carlin, and White (2009) proposed a jackknife procedure (Efron 1981) to estimate $\sqrt{B/M}$, the so-called Monte Carlo error. These authors then propose to select the sufficient number of imputed datasets based on achieving a predefined level of precision. Their approach is implemented in the command `mim` in STATA.

3. Number of Imputed Datasets: An Alternative Proposal

As we have seen, in most of the proposed rules for choosing M, the comparison is always between characteristics under a given, finite M on the one hand, and $M \to +\infty$ on the other. In contrast, our proposal is to compare the quantity of interest with its successor (i.e., under M vs. $M+1$). As the estimated parameter and its variance using MI will converge to some asymptotic values as $M \to \infty$, monitoring this convergence is insightful. This implies an iterative perspective, which is convenient for practice. We call this procedure the iterative multiple imputation (imi). The imi procedure is formulated as follows:

1. **Start.** Select an initial number of imputed datasets, M_0, $\widetilde{\boldsymbol{\theta}}_{M_0} = \sum_{i=1}^{M_0} \widehat{\boldsymbol{\theta}}_i / M0$.
2. **Update.** For $m > M_0$,

$$\widetilde{\boldsymbol{\theta}}_{m+1} = \frac{m\widetilde{\boldsymbol{\theta}}_m + \widehat{\boldsymbol{\theta}}_{m+1}}{m+1}. \tag{8}$$

3. **Distance.** Compute: $d_{m+1} = d(\widetilde{\boldsymbol{\theta}}_{m+1}, \widetilde{\boldsymbol{\theta}}_m)$ using an appropriate distance.
4. **Stopping rule.** $d_j < \varepsilon$ for $j = m+1, \ldots, m+k_0$.

Here, M_0 is an integer indicating the initial number of imputations. For $M_0 = 2$, the stopping rule will be examined from the beginning, but in situations where the user knows a minimum number of imputations are needed (based on proportion of missing data, etc.) a larger M_0 can be used. $\widehat{\boldsymbol{\theta}}_m$ is the estimated parameter in the mth iteration and $\widetilde{\boldsymbol{\theta}}_m = \sum_{i=1}^{M} \widehat{\boldsymbol{\theta}}_m / M$. Note

that $\widetilde{\boldsymbol{\theta}}$ can be replaced with other quantities of interest, for example, p-values. Clearly, for various quantities of interest, the corresponding combination rule should be applied in (8). Since the convergence of this procedure is not monotone, one needs to determine an integer k_0 as the number of successive steps that the stopping rule should be validated. This would prevent an early declaration of convergence. Again, $k_0 = 1$ would stop the first time this criterion is met, but a larger k_0 is recommended. Our proposal is to use $k_0 = 3$ (see Sections 4 and 6).

As can be seen, the number of imputed datasets in the proposed iterative procedure has two components: a deterministic part: $M_0 + k_0$ and a stochastic part which is determined by the iterations. Our implementation of this procedure in the R package `imi` makes it possible to determine these two parameters interactively after observing the computation time of imputing the dataset and fit the model for $M0 = 2$ times.

According to the convergence rate of the multivariate central limit theorem (Gotze 1991; Berckmoes, Lowen, and Van Casteren 2016), roughly speaking, the convergence of $\widetilde{\boldsymbol{\theta}}$ to θ as $M \to \infty$ in (4) depends on three different aspects: sample size, dimension of the random vector (parameter space), and the third moment of the components of the random vector. In our iterative procedure, the sample size consists of two aspects: the number of imputed datasets, M, also the proportion of missing data. A larger M makes the convergence faster, while a larger proportion of missing data makes it slower. On the other hand, another important issue is the dimension of the parameter space, the more parameters to estimate the more slowly the convergence will be achieved. The third moment highlights the role of the model, for example, the convergence for a linear model would be different from a logistic regression (see Section 4). Therefore, an appropriate method for determining M should consider all these aspects. In other words, it should be sensitive to changes in any of these factors. As our proposed method uses the goal of the analyses (parameter estimate, hypothesis testing, etc.) in the process of determining M, it includes such aspects. Simulation results in Section 4 support this claim as well.

One important aspect of the proposed iterative procedure is the choice of an appropriate distance in Step 3. One immediate choice is to use an Euclidean distance, as follows

$$d_{m+1}^{\text{Euc}} = \sqrt{\left(\widetilde{\boldsymbol{\theta}}_{m+1} - \widetilde{\boldsymbol{\theta}}_m\right)^T \left(\widetilde{\boldsymbol{\theta}}_{m+1} - \widetilde{\boldsymbol{\theta}}_m\right)}. \tag{9}$$

Alternatively, one may try to make the largest difference smaller, then an ℓ_∞-norm would be the appropriate distance. Obviously, for $q = 1$ these two are equivalent

$$d_{m+1}^{\ell_\infty} = \max\left\{|\widetilde{\boldsymbol{\theta}}_{m+1} - \widetilde{\boldsymbol{\theta}}_m|\right\}. \tag{10}$$

The common problem with ℓ_p-norm type distance measures in (9)–(10) is the fact that they would emphasize the elements in the parameter vector which have larger magnitude. Also, they are not robust against changes in units. This can also be seen from the general definition of the ℓ_p-norm of a vector $\mathbf{x} = (x_1, \ldots, x_k)$, $\|\mathbf{x}\|_p$

$$\|\mathbf{x}\|_p = \left(|x_1|^p + \ldots + |x_k|^p\right)^{1/p}, \quad p \leq 1.$$

When $p \to \infty$, one can show

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, \ldots, |x_k|\}.$$

As an alternative, we propose using the Mahalanobis distance (Mahalanobis 1936)

$$d_{m+1}^{\text{Mah}} = \sqrt{\left(\widetilde{\boldsymbol{\theta}}_{m+1} - \widetilde{\boldsymbol{\theta}}_m\right)^T S^{-1} \left(\widetilde{\boldsymbol{\theta}}_{m+1} - \widetilde{\boldsymbol{\theta}}_m\right)}. \tag{11}$$

For using (11), one needs to define an appropriate S. One can show

$$\widetilde{\boldsymbol{\theta}}_{m+1} - \widetilde{\boldsymbol{\theta}}_m = \frac{\widehat{\boldsymbol{\theta}}_{m+1} - \widetilde{\boldsymbol{\theta}}_m}{m+1}. \tag{12}$$

Considering the fact that $\widehat{\boldsymbol{\theta}}_{m+1}$ and $\widetilde{\boldsymbol{\theta}}_m$ are independent given the observed part of the sample, one may use

$$\begin{aligned} \text{cov}\left(\frac{\widehat{\boldsymbol{\theta}}_{m+1} - \widetilde{\boldsymbol{\theta}}_m}{m+1}\right) &= \frac{1}{(m+1)^2}\left[\text{var}(\widehat{\boldsymbol{\theta}}_{m+1}) + [\text{Var}(\widetilde{\boldsymbol{\theta}}_m)\right] \\ &= \frac{1}{(m+1)^2}[\text{var}(\widehat{\theta}_{m+1}) + \widehat{W}_m]. \end{aligned} \tag{13}$$

The division by $m + 1$ in (13) makes this an inappropriate candidate for S, as it changes every time, while we need to monitor convergence of the procedure, relative to a sufficiently stable yardstick.

Other choices for S include the within, between, or combined covariance matrix of the parameters at step $m + 1$, that is, $\widehat{W}_{m+1}$, $\widehat{B}_{m+1}$, or $\widehat{V}_{m+1}$, respectively; here $\widehat{V}$ is computed as in (5). Our simulation results show that $\widehat{B}_{m+1}$ cannot be an appropriate choice since being normalized by the variability of the differences, the resulting Mahalanobis distance is not sensitive to the fraction of missing data. This will be discussed in more detail in Section 4.

An appropriate distance should be sensitive to the fraction of missing data, but robust against rescaling model parameters. Both $\widehat{W}$ and $\widehat{V}$ serve these purposes, but our simulation results show that (see Section 4) for some large values of variance components, $\widehat{W}$ would show chaotic behavior from one iteration to another in some cases, while $\widehat{V}$ balances between and within variability, hence, behave more stably. Therefore, our proposal for an appropriate distance in Step 3 is the Mahalanobis distance with $S = \widehat{V}_{m+1}$. In cases where the quantity of interest is unitless, for example, a correlation coefficient, a coefficient of variation, a coefficient of determination, a p-value, etc., using ℓ_p-norm type distances such as in (9)–(10) is recommended, especially when obtaining the variance of the quantity of interest is not straightforward. Of course, one may use any tailored distance according to the problem at hand, this would not change the proposed iterative procedure.

As mentioned earlier, the convergence speed of the imi procedure depends on various factors such as the model, the parameter space, and the fraction of missing data. Also, the choice of the distance in Step 3 of the procedure plays a determining role when it comes to deciding when to stop. Therefore, formulating a fixed universal threshold for ϵ in Step 3 of the proposed iterative procedure seems not to be possible. However, based on extensive simulation results, when using a Mahalanobis-types distance as in (11) with $S = \widehat{V}_{m+1}$, considering $\epsilon = 0.05$ leads to a liberal choice of M. If one wants to select a conservative M, we propose to set $\epsilon = 0.01$. In the same setting, such choices

Table 1. Mean, SD for selected M given different ϵ and k_0 values, and number of times $M > 500$ (out of 100 replications) for different models and different ϵs using the Mahalanobis-type distance with $S = \widehat{V}$.

		$k_0 = 1$			$k_0 = 3$			$k_0 = 5$		
Model	ϵ	Mean	SD	$M > 500$	Mean	SD	$M > 500$	Mean	SD	$M > 500$
CS-10%	0.005	13.82	5.02	0.00	29.16	8.54	0.00	37.63	9.24	0.00
	0.01	7.15	3.08	0.00	15.38	4.88	0.00	19.31	6.01	0.00
	0.02	3.88	1.96	0.00	6.98	3.05	0.00	8.62	3.64	0.00
	0.03	2.67	1.44	0.00	4.36	2.22	0.00	5.31	2.51	0.00
	0.04	2.09	1.06	0.00	3.27	1.71	0.00	3.63	1.91	0.00
	0.05	1.68	0.84	0.00	2.42	1.51	0.00	2.78	1.80	0.00
CS-70%	0.005	30.71	10.28	0.00	71.12	13.25	0.00	94.98	14.39	0.00
	0.01	18.12	5.24	0.00	37.93	8.01	0.00	50.62	9.72	0.00
	0.02	10.06	3.74	0.00	20.73	5.33	0.00	25.04	5.41	0.00
	0.03	6.45	2.59	0.00	13.65	3.95	0.00	16.85	4.39	0.00
	0.04	4.81	2.10	0.00	9.61	3.03	0.00	12.24	3.78	0.00
	0.05	3.94	1.94	0.00	7.67	2.70	0.00	9.40	3.18	0.00
AR(1)-10%	0.005	18.86	7.24	0.00	39.05	11.13	0.00	46.60	12.93	0.00
	0.01	10.37	4.28	0.00	18.72	6.58	0.00	23.81	8.44	0.00
	0.02	4.99	2.33	0.00	9.85	3.57	0.00	11.74	4.29	0.00
	0.03	3.45	1.59	0.00	5.96	2.77	0.00	7.44	3.33	0.00
	0.04	2.40	1.22	0.00	4.15	2.29	0.00	4.67	2.54	0.00
	0.05	2.03	1.07	0.00	3.02	1.65	0.00	3.64	2.15	0.00
AR(1)-70%	0.005	44.30	13.06	0.00	97.12	17.89	0.00	123.94	20.28	0.00
	0.01	23.71	8.08	0.00	51.85	9.49	0.00	64.14	11.04	0.00
	0.02	13.55	4.93	0.00	26.79	6.21	0.00	33.91	6.82	0.00
	0.03	9.28	3.36	0.00	17.81	3.97	0.00	21.94	5.11	0.00
	0.04	6.76	2.65	0.00	13.59	3.34	0.00	16.67	3.93	0.00
	0.05	5.44	2.26	0.00	10.84	2.98	0.00	13.12	3.17	0.00
Logreg-10%	0.005	39.06	16.12	0.00	71.73	24.19	0.00	90.08	29.32	0.00
	0.01	19.95	8.71	0.00	38.31	14.02	0.00	46.67	15.70	0.00
	0.02	11.14	5.26	0.00	20.38	7.42	0.00	25.47	9.62	0.00
	0.03	7.31	3.60	0.00	13.06	5.50	0.00	16.28	6.07	0.00
	0.04	5.36	2.82	0.00	9.98	4.10	0.00	11.72	4.71	0.00
	0.05	4.60	2.56	0.00	7.85	3.38	0.00	9.53	3.78	0.00
Logreg-70%	0.005	61.33	26.34	0.00	140.31	53.29	0.00	174.82	67.77	0.00
	0.01	36.68	15.25	0.00	76.05	25.85	0.00	94.65	32.48	0.00
	0.02	21.35	7.88	0.00	40.75	13.77	0.00	51.16	16.31	0.00
	0.03	14.68	6.37	0.00	28.97	8.96	0.00	34.92	10.65	0.00
	0.04	11.76	4.29	0.00	22.35	6.44	0.00	26.23	7.47	0.00
	0.05	10.01	4.00	0.00	18.99	5.44	0.00	21.77	6.26	0.00

NOTE: The results are presented for three different successive validation steps $k_0 = 1, 3, 5$.

of ϵ would select Ms that are approximately in agreement with Rubin's classic suggestion, for example, to use $M = 5$ for 10% missing data (see Table 1).

In general, for selecting ϵ, one needs to consider the nature of the quantity of interest when deciding on ϵ. For example, when dealing with p-values, one needs to ensure that increasing the number of imputed datasets would not change the result of the test, so the distance between two p-values from imputed datasets number m and $m + 1$, near or smaller than the significance, α, should be smaller than, for example, $\alpha/10$; see Section 5.1. Therefore, in case of p-values, we propose to use a Euclidean distance with $\epsilon = \alpha/10$. Note that, given the null hypothesis, the p-value is uniformly distributed. Therefore, for an appropriate ϵ and up to a constant coefficient, using a Mahalanobis distance is equivalent to the Euclidean distance.

As Wald (1939) pointed out, the two central procedures in theory of statistics are parameter estimation and hypothesis testing. To summarize our discussion above, we proposed to use a different ϵ for each of these procedures. In case of parameter estimation one may use 0.05 or 0.01, and in case of hypothesis testing (using p-values) one may use $\alpha/10$.

4. Simulation Study

To study our proposed method, two simulation plans are considered. One with a multivariate normal vector and the other with a logistic regression model. Using these two simulation settings, the performance of our proposed procedure for both parameter estimation and hypothesis testing will be studied. The results will be displayed, compared, and discussed, after presenting the plans.

4.1. Simulation Plan

A simulation study is undertaken for a multivariate normal vector as well as data generated from a logistic regression model. Two types of covariance structures are considered for the multivariate normal model: compound-symmetry (CS) and first-order autoregressive (AR(1)). Consider $\boldsymbol{Y}_i$ a vector of length k, then $\boldsymbol{Y}_i \sim N(\mu \mathbf{1}_k, \Sigma)$. Under CS, $\Sigma = \sigma^2 I_k + \tau J_k$, for $\sigma^2, \tau > 0$ and I_k and J_k the identity and all-ones matrices, respectively. For AR(1), we have $\Sigma = \sigma^2 C$ where the (i, j) element of C is defined as $\rho^{|i-j|}$, for $\sigma^2 > 0, -1 \leq \rho \leq +1$.

All data are generated for $\mu = 0$. The length of each random vector is set to 5, and the sample size is considered to be 100. For each covariance structure in the multivariate normal case 18 different scenarios are considered as combinations of the following settings:

- Two proportions of missing data are used: 10% and 70%;
- three σ^2 values are used: 1, 16, 64;
- three ρ values are used: 0.2, 0.5, 0.8. In case of CS, the corresponding τ is computed given the specified ρ using $\tau = \rho\sigma^2/(1-\rho)$.

Logistic regression data are generated with parameters $(\beta_0, \beta_1, \beta_2) = (0.2, -2, 0.5)$, where the design matrix is generated using each of the nine considered CS settings. The parameters of the CS and AR(1) models are estimated using the results in Hermans et al. (2019a) and Hermans et al. (2019b), respectively. The logistic regression is fitted using R function `glm`. The proportion of missing values is the same as in the multivariate normal scenarios (10% and 70%).

The missing data for all of these scenarios are generated under a missing at random (MAR) mechanism using the function `ampute` in the R package `mice` (see Schouten et al. 2017; Schouten, Lugtig, and Vink 2018). The function `ampute` implements a multivariate approach to generate missing data by assigning a weight to each observation. In case of a MAR mechanism, these weights only depend on the observed part of the sample. Based on the assigned weights, a logistic distribution will be used to compute the probability of missingness. An observation with a larger weight has more chance to be missing.

Each incomplete dataset is imputed 500 times using a multivariate normal predictive model in the R package `Amelia2`, and each scenario is repeated 100 times as well.

To compute the distance between two steps in our iterative procedure, we use five different measures: Euclidean distance (9), ℓ_∞-norm (10), Mahalanobis distance (11) with S selected as within ($\widehat{W}$), between ($\widehat{B}$), and combined covariance matrix ($\widehat{V}$) of the estimated parameter vector.

There are two main outcomes of interest from different considered scenarios. First, the convergence plot for different distances. Second, the M for which the iterative procedure terminates. Such M is computed using seven values $\epsilon = 0.005, 0.01, 0.02, 0.03, 0.04$, and 0.05, and for each ϵ, three successive validation steps are considered: $k_0 = 1, 3$, and 5.

To evaluate our proposed method in case of p-values, one-sample t-tests are applied with $H_0 : \mu = d_0$ on the first column of the data generated from each setting of CS and AR(1) covariance structure. In addition, paired t-tests are also performed for testing $H_0 : \mu_1 - \mu_2 = d_0$ for the first and the second columns of data generated using each setting of CS and AR(1) covariance matrices. To see the effect of the test value on the convergence, we have considered $d_0 = 0$ and $d_0 = 10$. As the p-value is a scalar, we use Euclidean distance (9) to measure distance of two successive steps. In case of p-values, again the convergence plots are provided; also, the sufficient number of imputed datasets is computed for two values of $\epsilon = 0.05/10$ and $0.05/100$. For each ϵ, three successive validation steps are considered: $k_0 = 1, 3$, and 5.

4.2. Simulation Results

As displaying the results of 18 different scenarios takes a lot of space, here we discuss all of the outcomes but only display the results for one of them. The rest of the results will be available in the R package `imi` via nine lists of datasets, the function `imi.make.plots` can be used to produce similar plots as in Figure 1 and Figure 2 for other scenarios. Further information can be found in the documentation of the package.

Figure 1 shows the convergence plots for the scenario with $\sigma^2 = 16$ and $\rho = 0.5$. As one may see, while using all of the distance functions a decreasing trend can be observed, but as expected Euclidean and ℓ_∞ norms, as well as Mahalanobis distance with $S = \widehat{B}$ are not performing well. On the other hand, the latter with both $S = \widehat{W}$ and $S = \widehat{V}$ performs fine. Note that for 70% missing data for some of the imputed datasets, quasi-complete separation occurs when fitting logistic regression. In such cases, a new imputed dataset is generated.

An interesting observation from Figure 1 is the different convergence rate for logistic regression compared with the multivariate normal case. As one may see, for 10% missing data, the logistic regression converges almost with the same speed as 70% missing data in case of a multivariate normal vector parameter estimates. That highlights the effect of the model (thus not only the proportion of missing data) when determining a sufficient number of imputed datasets.

Figure 2 shows the convergence plots using $M = 500$ for the p-values of one sample and paired t-tests for two different test values (0, 10). The distance between two p-values is computed using the Euclidean distance. As one may see, one important factor that should be taken into account when determining M is the test value. When it is far from the actual parameter convergence is achieved for a much smaller M.

To explore the selected M using various ϵs and k_0s to evaluate our proposals ($\epsilon = 0.01$ or 0.05, and $k_0 = 3$), Tables 1 and 2 show the selected M for parameter estimation in multivariate normal and logistic regression (Table 1) and p-values of one sample and paired t-tests (Table 2). As one may see, again the effect of proportion of missing data and the model is verified on the selected sufficient M. While a dataset with a larger proportion of missing data needs a larger M, the number of sufficient imputed datasets in case of logistic regression is different from multivariate normal, even for a smaller proportion of missing data. This would highlight the role of a procedure that accounts for all of these aspects.

Also, it seems that our proposals for ϵ and k_0 are sensible in terms of comparing with Rubin's classical rule suggesting 3–5 imputed datasets for 10% missing data. In case of p-values it seems $\epsilon = 0.005$ together with $k_0 = 3$ works acceptably fine.

An interesting observation when comparing the selected M for different values of ρ is the effect of *missing information*. For the same fraction of missing data, when the correlation is larger, the selected M is smaller. That would suggest that what really matters is not only the fraction of missing data, but the fraction of missing information. As an extreme case, consider two variables X and Y in a given dataset, suppose X is complete and Y is partially missing. Assume further that $Y_i = X_i$ ($\rho = 1$), for every subject. In this case one imputation would be sufficient, no matter the fraction of missing items. Figure 3 shows the

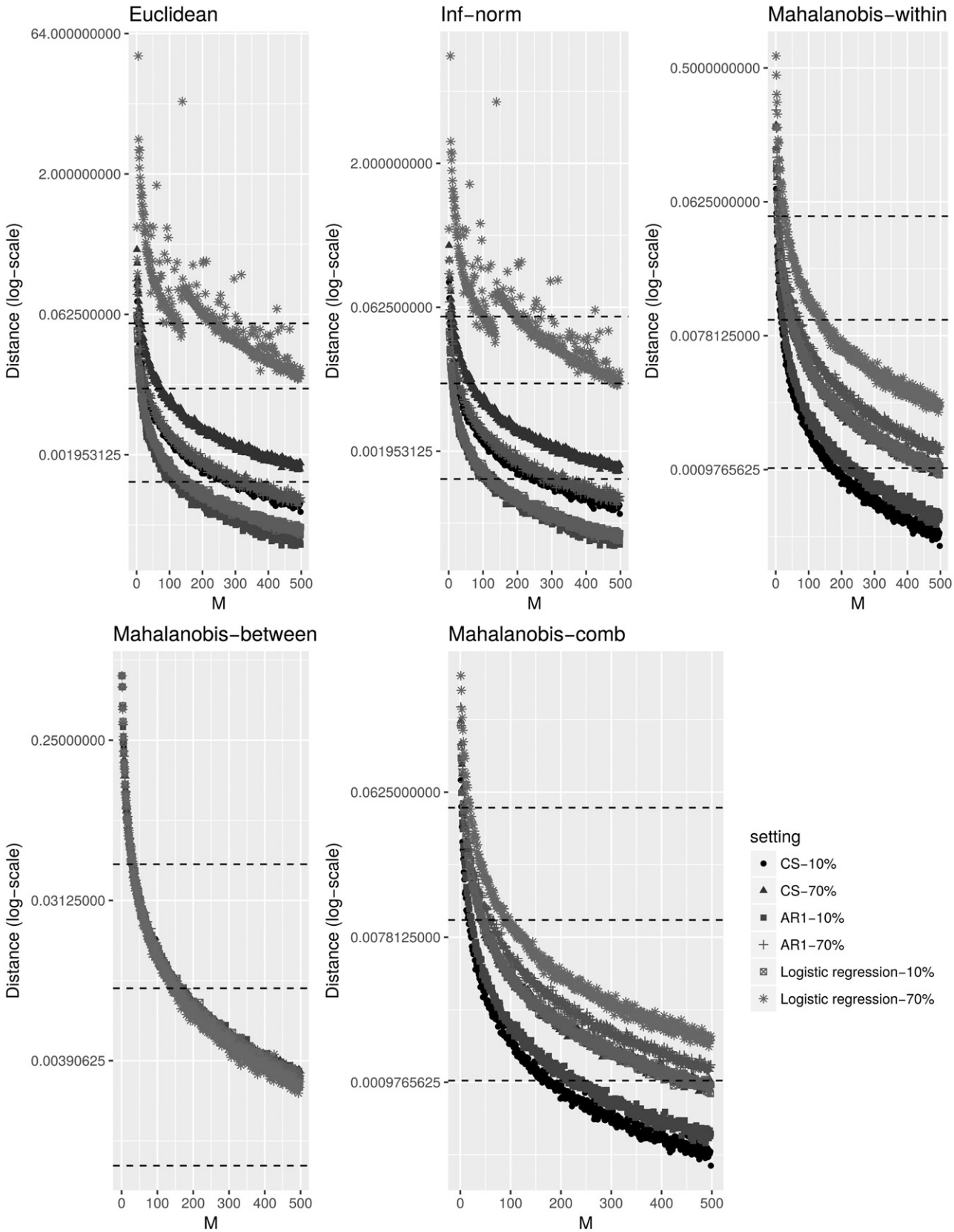

Figure 1. Convergence plots for multivariate normal random vectors parameter estimates with CS and AR(1) covariance matrix structures (with $\sigma^2 = 16$ and $\rho = 0.5$), as well as a logistic regression model, using five different distance functions: Euclidean norm, ℓ_∞-norm, and Mahalanobis distance with within, between, and combined covariance matrices over imputed sets of data for two proportion of missing data: 10% and 70%. Each point in the plot is the average over 100 replications. The three horizontal dashed lines are $\epsilon = 0.05, 0.01, 0.001$, respectively.

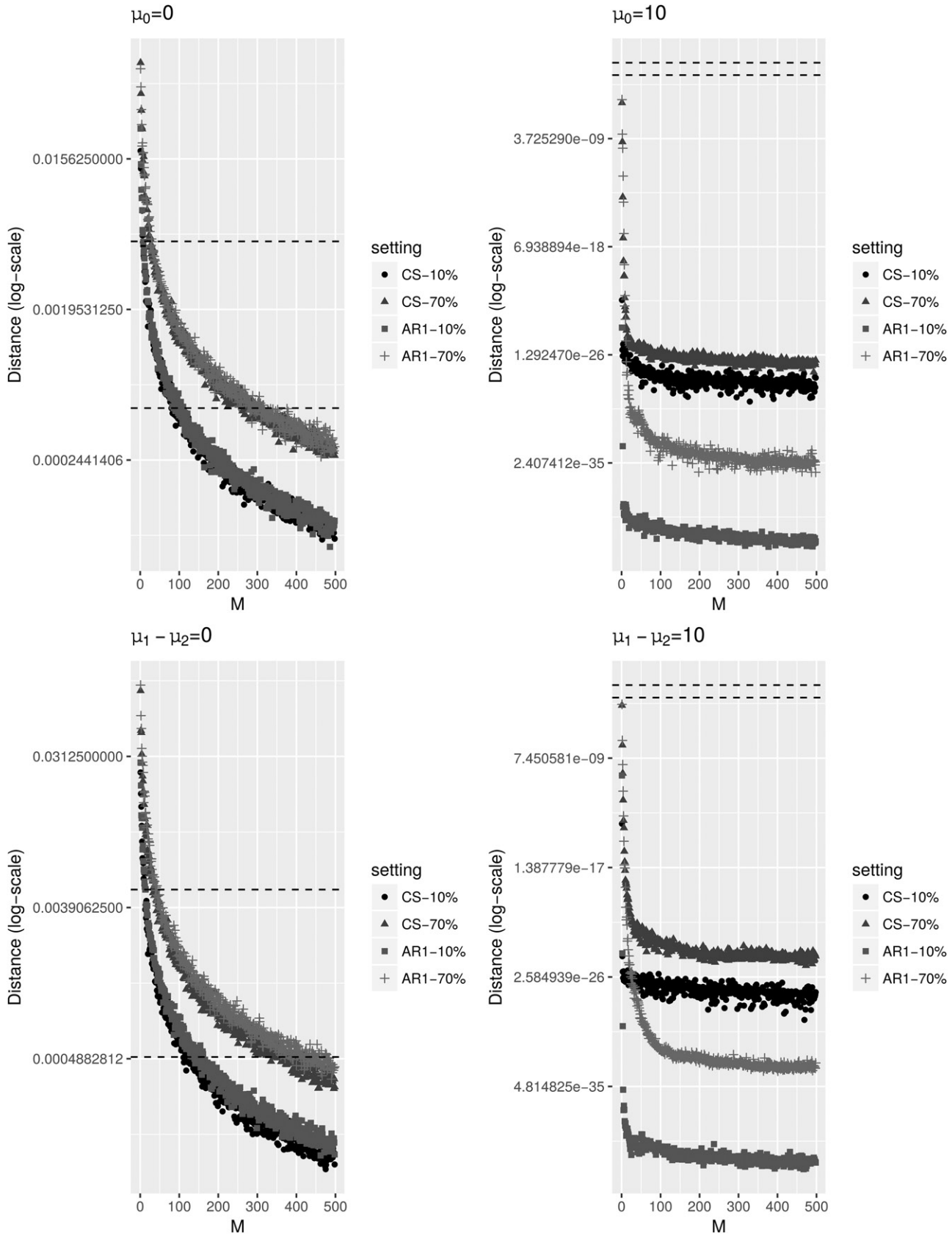

Figure 2. Convergence plots for the *p*-values of one sample (first row) and paired (second row) *t*-tests for different values of test statistics. The data are generated from a (multivariate) normal distribution with mean 0 and CS and AR(1) covariance matrices (with $\sigma^2 = 16$ and $\rho = 0.5$). The missing values are generated with proportions 10% and 70%. The two horizontal dashed lines are $\epsilon = 0.005, 0.0005$, respectively.

Table 2. Mean, SD for selected M given different ϵ and k_0 values, and number of times $M > 500$ for different models and different ϵs for t-test and paired t-test with test value $\mu = 0$.

			$k_0 = 1$			$k_0 = 3$			$k_0 = 5$		
Test	Model	ϵ	Mean	SD	$M > 500$	Mean	SD	$M > 500$	Mean	SD	$M > 500$
$\mu_1 = 0$	CS-10%	0.005	2.90	2.55	0.00	8.17	6.74	0.00	11.32	8.93	0.00
		5e−04	12.78	9.70	0.00	50.41	32.28	0.00	83.97	53.85	0.00
	CS-70%	0.005	5.41	3.85	0.00	20.22	10.08	0.00	29.20	14.79	0.00
		5e−04	21.75	12.38	0.00	119.70	57.55	0.00	211.70	93.22	0.00
	AR(1)-10%	0.005	2.76	2.03	0.00	7.55	6.39	0.00	10.55	9.08	0.00
		5e−04	12.73	10.51	0.00	53.61	34.37	0.00	88.68	63.04	0.00
	AR(1)-70%	0.005	5.91	4.34	0.00	20.98	10.62	0.00	30.73	15.85	0.00
		5e−04	22.88	15.28	0.00	120.90	57.80	1.00	214.83	97.30	1.00
$\mu_1 - \mu_2 = 0$	CS-10%	0.005	3.32	2.24	0.00	10.34	7.32	0.00	15.16	11.30	0.00
		5e−04	15.75	11.31	0.00	64.97	37.27	0.00	110.12	58.83	0.00
	CS-70%	0.005	6.62	4.14	0.00	25.41	13.59	0.00	36.49	19.45	0.00
		5e−04	25.94	17.69	0.00	157.88	74.58	6.00	243.37	124.17	6.00
	AR(1)-10%	0.005	4.02	2.78	0.00	11.46	7.58	0.00	16.18	10.01	0.00
		5e−04	14.37	10.65	0.00	74.38	40.09	0.00	134.99	70.34	0.00
	AR(1)-70%	0.005	8.31	5.11	0.00	29.14	14.83	0.00	43.30	19.48	0.00
		5e−04	31.39	17.23	0.00	160.09	72.00	2.00	288.81	112.58	2.00

NOTE: The results are presented for three different successive validation steps $k_0 = 1, 3, 5$.

selected M for different scenarios with $\sigma^2 = 16$ and $\rho = 0.5$ and 0.8, and $k_0 = 3$. As one may see, in most of the cases, for the same proportion of missing data and model, the selected M is smaller when ρ is larger.

5. Applications

In this section, two applications of MIs are considered: fitting a logistic regression to incomplete data, as well as combining p-values from a one-sample t-test for a set of incomplete data. The number of imputed datasets will be determined using the proposed iterative procedure via the R package `imi`.

5.1. Leuven Eye Study: Logistic Regression

The Leuven Eye Study (LES; Pinto et al. 2015), is an extensive observational study of glaucoma performed at the ophthalmology department of UZ Leuven. The dataset so far consists of 141 variables measured for 585 subjects. As this study was performed in a clinical setting, it was not feasible to measure all these variables for all of the subjects. Because of that, missingness is a considerable problem. Among these 141 variables, 130 of them are selected to be used in MI. An analysis was performed to study which risk factors are relevant for the binary outcome defined as normal versus glaucoma. The risk factors selected for the model include both structural and functional measurements of the eye (cup to disc ratio, corneal thickness, visual acuity), previous medical history (gender, sleep apnea, rhythm disorder, having had cataract surgery, being medicated with statins or calcium channel blockers) and more variable biometric measurements, like intraocular pressure and diastolic blood pressure.

Figure 4 (bottom-left) shows a histogram of the number of missing values (out of 585 patients) in the LES dataset. As one may see, the number of missing values is diverse. Also the patterns of missing values could be different from variable to variable. Therefore, determining M using classical or simulation-based approaches could be a challenge here. Using our proposal, it could be more convenient to determine the sufficient number of imputed datasets.

Considering the size of the dataset as well as various variable types, to impute the missing values the fully conditional specification (FCS) approach (van Buuren 2007, 2012; van Buuren et al. 2006) is employed. For continuous variables, predictive mean matching (Little 1988) is used for generating the imputed datasets. Also, a binary logistic regression predictive model is used to impute binary variables. In the specific case of one variable with three levels, a polytomous logistic regression predictive model is used. To create a sufficient number of imputed datasets we use the function `imi.glm` in the R package `imi`.

We have selected 2 for the initial number of imputations (M_0). Also, the convergence criterion should be successively validated 3 times to terminate the procedure ($k_0 = 3$). Figure 4 (bottom-right) shows the convergence plot for this model on the LES data. As one may see, with $\epsilon = 0.05$ (the liberal choice) and $k_0 = 3$ we need $M = 58$ imputed datasets, while with the conservative choice, $\epsilon = 0.01$, one may need to generate $M = 250$ imputed datasets.

5.2. Cholesterol Data: One Sample t-Test p-Values

The cholesterol dataset includes cholesterol levels for 28 patients treated at a Pennsylvania medical center. The cholesterol levels have been recorded on day 2, day 4, and day 14 after an attack for each patient. However, there are nine missing values for day 14. This dataset was analyzed by Schafer (1997, chap. 5) and is publicly available in R package `norm2`.

Here we use our R implemented function `imi.t.test` to sufficiently impute this dataset and perform the t-test on it for two different test-values ($\mu_0 = 200, 220$).

Again, we ask for two initial imputations and then three successive validation steps ($M_0 = 2, k_0 = 3$). Figure 4 (top) shows the convergence plot for $\mu_0 = 200$ (left) and $\mu_0 = 220$ (right). In each case the vertical dashed line shows the selected M for different values of ϵ. As one may see, when testing $H_0 : \mu = 200$, for $\epsilon = 0.05/10$ and $k_0 = 3$, only $M = 9$ imputed datasets are sufficient, changing ϵ to $0.05/100$ will increase M to

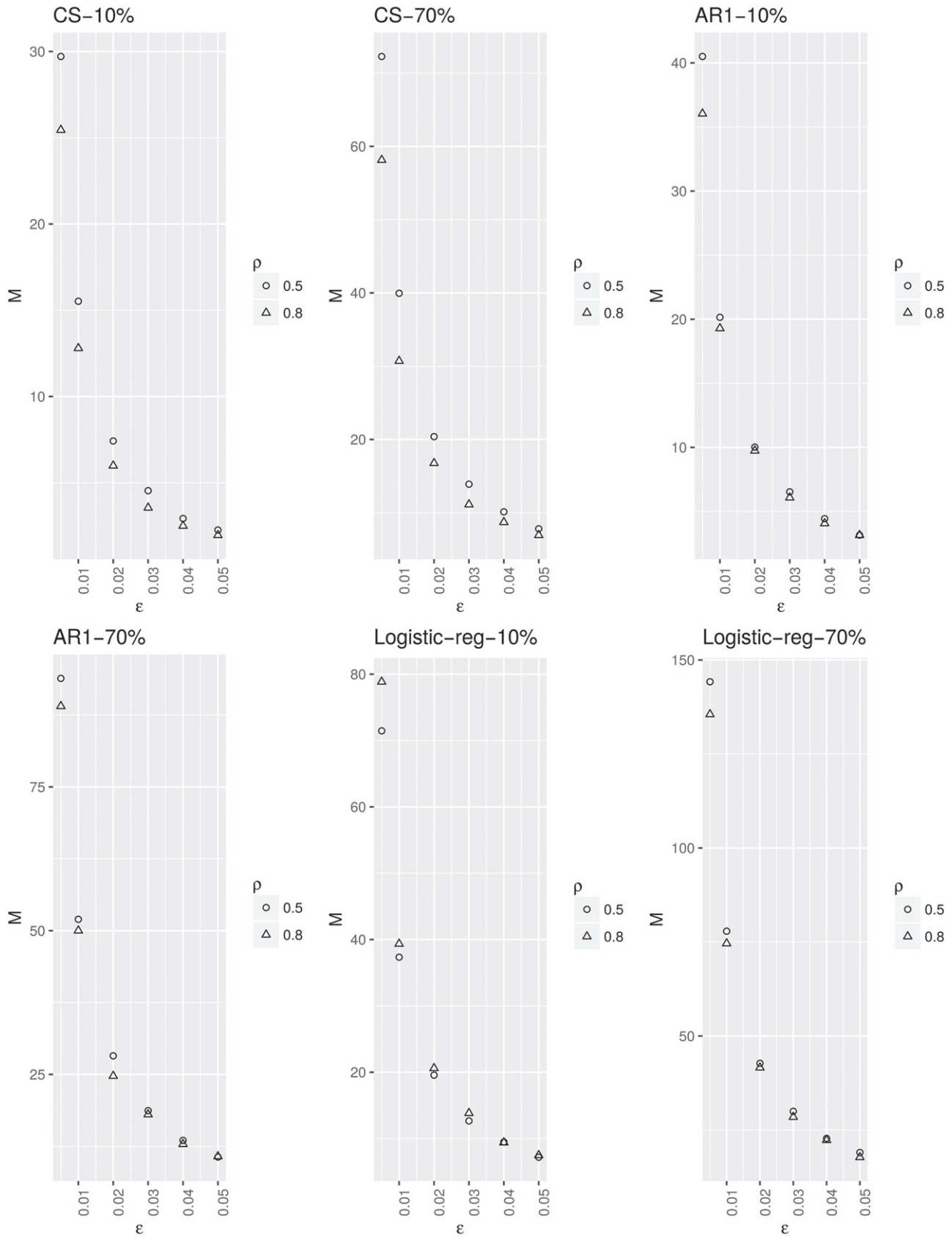

Figure 3. Averaged selected M with $k_0 = 3$ for data generated from a multivariate normal with mean 0 and covariance matrix with CS and AR(1) structures with $\sigma^2 = 16$ and $\rho = 0.5, 0.8$, as well as a logistic regression model. The proportions of missing data are 10% and 70%. Each point in the plot is the average over 100 replications.

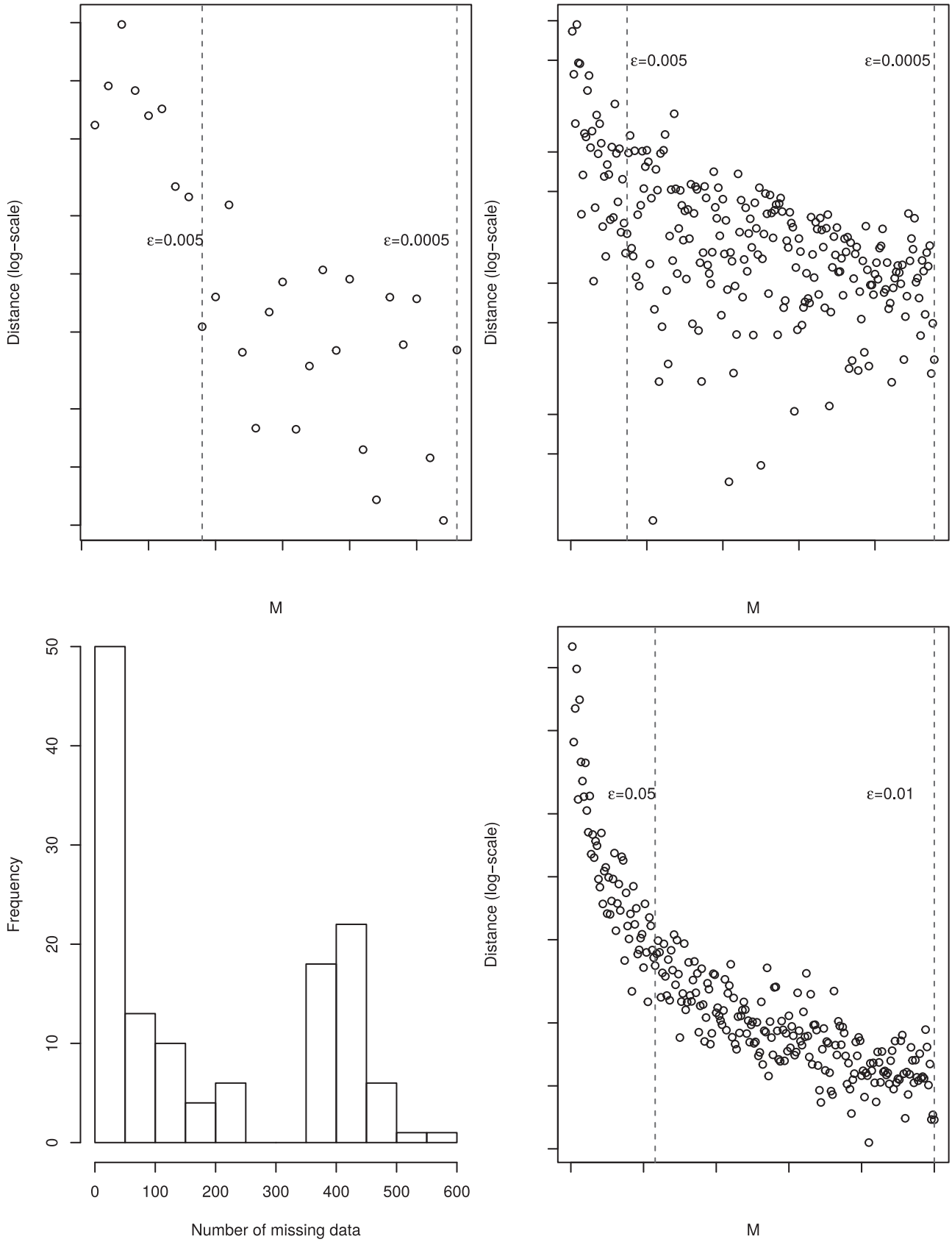

Figure 4. Convergence plots for *p*-values of *t*-test on cholesterol data (top) with $\mu_0 = 200$ (left) and $\mu_0 = 220$ (right) test values, and LES logistic regression (bottom-right), with $\epsilon = 0.05, 0.01$, $M_0 = 2$, and $k_0 = 3$. The dashed vertical line shows the selected M for different values of ϵ. The histogram for the number of missing values in LES dataset (bottom-left) is also presented.

28. When testing $H_0 : \mu = 220$, however, more imputations are needed. For $\epsilon = 0.05/10$ we get $M = 37$ and for $\epsilon = 0.05/100$ we obtain $M = 239$.

6. Discussion and Concluding Remarks

MI is an appealing and extensively used method to analyze incomplete data, or, even more broadly, data that can be cast in a missing-data framework (Van der Elst et al. 2015). There is widely held and largely justified wisdom that a small number of imputation suffices for many practical purposes, even though Schafer (1997) stated that the number depends on the amount of missing information, the data type under study, and the inferential purpose. We have presented a very simple procedure, that is based on conventional large-sample convergence results, by merely using the fact that MI is in itself a sampling mechanism. It is based on comparing the estimates under M and $M + 1$ imputations, the procedure stops when the distance between two steps become smaller than a predefined ϵ. One needs to select an initial number of imputations as well as number of steps the stopping rule should be validated. This would prevent early stopping when two consecutive quantities of interest are unusually close. Thus, when convergence is suggested, to play safe, one could still go on for a while, and wait until convergence has been confirmed. Stopping too early would lead to less precise estimates, or unwanted decision making, while the main problem with too high an M is the computational cost, also when the distance between two steps becomes very small, due to a small ϵ (or equivalently selecting a large M), some numerical underflow issues could occur.

There are two main parameters that are needed to be specified before applying our procedure on an incomplete dataset. The stopping rule for the distance between two steps, ϵ, and the number of steps this criterion should be successively validated, k_0. Based on our numerical experiences, we suggest $k_0 = 3$ validation steps is enough in most of the applications. For ϵ, the choice depends on the chosen distance which itself depends on the purpose of the analysis: estimation or hypothesis testing. For estimation purposes, that is, estimating the parameters and their precision, we suggest to use a Mahalanobis distance with $S = \widehat{V}_{M+1}$, then $\epsilon = 0.05$ as a liberal choice and $\epsilon = 0.01$ as a conservative choice. For p-values at $100(1 - \alpha)\%$ level of confidence, we suggest to use the Euclidean distance with $\alpha/10$ as a liberal choice, and as a conservative choice our proposal is to use $\alpha/100$. Of course, as discussed in Section 3, the convergence rate depends on various aspects (proportion of missing data, the model, dimension of the parameter space). Therefore, it is very well possible that for less conventional applications one needs different tailor-made values for ϵ and k_0.

Simulations and case studies, the latter with admittedly a large fraction of missingness, indicate that it might be needed, in many practical settings, to generate a number of imputations well above what common wisdom prescribes.

Our method has several advantages. First, it is applicable in any context where MI is useful, irrespective of the amount of missing information, the model used, or the target of inference. Second, imputed datasets can be generated one by one, and each time one can easily decide about the need to continue adding new imputed datasets, depending on the research question(s) to be answered. Third, the procedure can easily be automated since "expert judgment" is not needed. Finally, no post-hoc sensitivity assessment to the number of imputations is required, allowing the stopping criterion to be specified prior to the data analysis or even prior to the data collection.

Furthermore, our implementation of this procedure in R package `imi` could make it more available for different applications. The current version of this package includes the t-test (one sample, two samples, and paired), as well as linear and generalized linear regression models. For each application one has the possibility to generate a sufficient number of imputed datasets. In case of existence of some already generated imputed sets of data, one can examine their sufficiency for the desired analyses. When they are not sufficient it is also possible to generate new imputed datasets till M becomes sufficiently large. The considered interactive options allow the user to select M_0 and/or k_0 based on the time it takes to generate two imputed datasets and perform the analyses on them.

Acknowledgments

We are grateful for suggestions made by anonymous referees, which have greatly helped to improve this article. We also wish to thank Ophthalmology Department of UZ Leuven for providing the Leuven Eye Study dataset.

Funding

The authors gratefully acknowledge the financial support from the IAP research network # P7/06 of the Belgian Government (Belgian Science Policy). The research leading to these results has also received funding from the European Seventh Framework Programme FP7 2007–2013 under grant agreement no. 602552. We gratefully acknowledge support from the IWT-SBO ExaScience grant.

References

Berckmoes, B., Lowen, R., and Van Casteren, J. (2016), "Stein's Method and a Quantitative Lindeberg CLT for the Fourier Transforms of Random Vectors," *Journal of Mathematical Analysis and Applications*, 433, 1441–1458. [3]

Bodner, T. E. (2008), "What Improves With Increased Missing Data Imputations?," *Structural Equation Modeling*, 15, 651–675. [2]

Carpenter, J. R., and Kenward, M. G. (2008), "Missing Data in Clinical Trials: Practical Guide," Birmingham: National Institute for Health Research, Publication RM03/JH17/MK. [1,2]

——— (2012), *Multiple Imputation and Its Application*, Chichester: Wiley. [1,2]

Efron, B. (1981), "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," *Biometrika*, 68, 589–599. [2]

Gotze, F. (1991), "On the Rate of Convergence in the Multivariate CLT," *The Annals of Probability*, 19, 724–739. [3]

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007), "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory," *Prevention Science*, 8, 206–213. [2]

Harel, O., and Schafer, J. (2003), "Multiple Imputation in Two Stages," in *Proceedings of Federal Committee on Statistical Methodology 2003 Conference*. [2]

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M. G., Van der Elst, W., Aerts, M., and Verbeke, G. (2019a), "Clusters With Unequal Size:

Maximum Likelihood Versus Weighted Estimation in Large Samples," *Statistica Sinica* (forthcoming). [5]

——— (2019b), "Fast, Closed-Form, and Efficient Estimators for Hierarchical Models With AR(1) Covariance and Unequal Cluster Sizes," *Communications in Statistics-Simulation and Computation*, 47(5), 1492–1505. [5]

Little, R. J. (1988), "Missing-Data Adjustments in Large Surveys," *Journal of Business & Economic Statistics*, 6, 287–296. [8]

Lu, K. (2017), "Number of Imputations Needed to Stabilize Estimated Treatment Difference in Longitudinal Data Analysis," *Statistical Methods in Medical Research*, 26, 674–690. [2]

Mahalanobis, P. C. (1936), "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Sciences*, 2, 49–55. [3]

Pinto, L. A., Willekens, K., Van Keer, K., Shibesh, A., Vandewalle, E., Molenberghs, G., and Stalmans, I. (2015), "Leuven Eye Study—Baseline and Methods," *Acta Ophthalmologica*, 93. [8]

Royston, P. (2004), "Multiple Imputation of Missing Values," *Stata Journal*, 4, 227–241. [2]

Royston, P., Carlin, J. B., and White, I. R. (2009), "Multiple Imputation of Missing Values: New Features for mim," *Stata Journal*, 9, 252. [2]

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Survey Research Methods Section of the American Statistical Association* (Vol. 1), American Statistical Association, pp. 20–34. [1]

——— (1979), "Illustrating the Use of Multiple Imputations to Handle Nonresponse in Sample Surveys," *Bulletin of the International Statistical Institute*, 48, 517–532. [1]

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley. [1,2]

Rubin, D., and Little, R. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [1,2]

Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374. [2]

Schafer, J. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall. [1,8,11]

Schouten, R., Lugtig, P., Brand, J., and Vink, G. (2017), "Multivariate Amputation Using Ampute," available at *https://cran.r-project.org/web/packages/mice/vignettes/ampute.html*. [5]

Schouten, R. M., Lugtig, P., and Vink, G. (2018), "Generating Missing Values for Simulation Purposes: A Multivariate Amputation Procedure," *Journal of Statistical Computation and Simulation*, 88, 2909–2930. [5]

van Buuren, S. (2007), "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification," *Statistical Methods in Medical Research*, 16, 219–242. [8]

——— (2012), *Flexible Imputation of Missing Data*, Boca Raton, FL: Chapman & Hall/CRC. [1,8]

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., and Rubin, D. B. (2006), "Fully Conditional Specification in Multivariate Imputation," *Journal of Statistical Computation and Simulation*, 76, 1049–1064. [8]

Van der Elst, W., Hermans, L., Verbeke, G., Kenward, M., Nassiri, V., and Molenberghs, G. (2015), "Unbalanced Cluster Sizes and Rates of Convergence in Mixed-Effects Models for Clustered Data," *Journal of Statistical Computation and Simulation*, 86, 1–17. [1,11]

Wagstaff, D. A., and Harel, O. (2011), "A Closer Examination of Three Small-Sample Approximations to the Multiple-Imputation Degrees of Freedom," *Stata Journal*, 11, 403–419. [2]

Wald, A. (1939), "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *The Annals of Mathematical Statistics*, 10, 299–326. [4]

White, I. R., Royston, P., and Wood, A. M. (2011), "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice," *Statistics in Medicine*, 30, 377–399. [2]