Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

A Note on Convergence Diagnostics for Multiple Imputation using Chained Equations (MICE)

Hanne Oberman
Utrecht University

Abstract

This Research Report contains a simulation study that serves as the basis of the technical paper that will be submitted for publication in *Journal of Statistical Software*. I have chosen to use the first format from the Research Report guidelines: "*It is written as a (mini) thesis, with an introduction, methods section, some results (i.e., preliminary analyses, or pilot simulations), and a discussion of results. The length of the research report should be maximally 2500 words of text (without references list and or tables and figures). Please do not include appendices, and no more than 6 tables or figures. Table and Figure captions do not count towards the word limit. An abstract may be included, but is not necessary*". I aim to publish a pre-print of this research report on ArXiv (`https://arxiv.org/`). That way, I can refer to the simulation study described here, without 'bulking up' the technical paper that I want to submit for publication in JSS. Another option would be to attach this note as online appendix to the technical paper on ShinyMICE.

Keywords: multiple imputation, convergence, **mice**, R.

1. Introduction

At some point, any scientist conducting statistical analyses will run into a missing data problem (Allison 2001). Missingness is problematic because statistical inference cannot be performed on incomplete data, and ad hoc solutions (e.g., listwise deletion) can yield wildly invalid results (Van Buuren 2018). To circumvent the ubiquitous problem of missing information, Rubin (1987) proposed the framework of multiple imputation (MI). MI is an iterative algorithmic procedure in which missing data points are 'guessed' (i.e. imputed) several times. [Add some stuff here about MICE.] The variability between the imputations validly reflects how much uncertainty in the inference is due to missing information–that is, if all statistical assumptions are met (Rubin 1987) [adjust this sentence to be more clear!].

[Feedback: The order of the introduction is not intuitive. some terms are mentioned first and only explained later. There are some short sentences with a lot of content, so implicitly a lot happens. Expand these sentences or delete (because the thing would be too obvious?). Explain 'true data generating model'. Follow up other important concepts with explanation or examples too.]

With MI, many assumptions are made about the nature of the observed and missing parts of the data and their relation to the 'true' *data generating model* (Van Buuren 2018) [explain what this is!]. Without proper evaluation of the imputations and the underlying assumptions, any drawn inference may erroneously be deemed valid. Such evaluation measures are currently missing or under-developed in MI software, like the world leading R package **mice** (Van Buuren and Groothuis-Oudshoorn 2011).

[Feedback: Look at abbreviations. For example r hat is just metnioned as 'e.g.'. It's not clear in the introduction section how important it is later.]

A fundamental assumption of MICE is convergence of the algorithm. MICE is a sort of Gibbs sampler, a type of Markov chain Monte Carlo (MCMC) algorithm. In general, the validity of inference resulting from MCMC algorithms is threatened by non-convergence. Hence we use convergence diagnostics to flag non-convergence. But it is not known whether conventional convergence diagnostics for MCMC methods work on MI data.

Conventional thresholds to diagnose non-convergence– e.g., Gelman and Rubin's 1992 statistic $\widehat{R} < 1.1$ –are not applicable on multiply imputed data (Lacerda, Ardington, and Leibbrandt 2007) [Expand or remove Rhat here! Refer to future section?]. Therefore, empirical researchers have to rely on visual inspection procedures that are theoretically equivalent to $\widehat{R}$ (White, Royston, and Wood 2011) [explain why this is appropriate, and numerical is not]. Visually assessing convergence is not only difficult to the untrained eye, it might also be futile. The convergence properties of MI algorithms lack scientific consensus (Takahashi 2017), and some default MICE techniques might not converge to stable distributions at all (Murray 2018). Moreover, convergence diagnostics for MI methods have not been systematically studied (Van Buuren 2018).

This paper investigates convergence properties of MI algorithms by means of model-based simulation in R (Core Team 2017). All programming code used in this note is available in the file 'XYZ.R' along with the manuscript, and on Github repository 'XYZ'. The results of this simulation study are guidelines for applied researchers to evaluate convergence of MI algorithms.

1.1. Terminology

The aim of this paper is to provide applied researchers a tool to assess convergence of multiple imputation algorithms. The intended audience consists of empirical researchers and statisticians who use multiple imputation to solve missing data problems. Basic familiarity with multiple imputation methodology is assumed. For an accessible and comprehensive introduction to MI from an applied perspective, see Van Buuren (2018). For the theoretical foundation of MI, see Rubin (1987).

[explain difference MI, MICE, mice]

[Explain the use of 'imputation chain' (the iterations of an imputation, only the final imputed values are used), starting values of chains, 'simulation condition' (the number of iterations

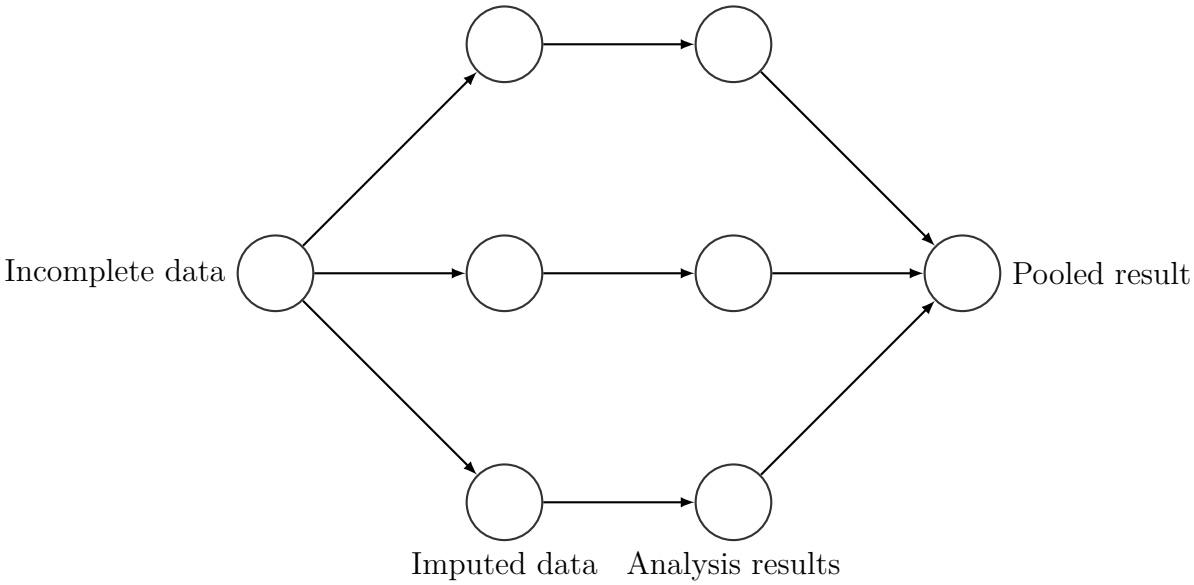

Figure 1: Scheme of main steps in multiple imputation (Adapted from (Van Buuren 2018, § 1.4.1))

the algorithm has before imputing), ...?]

[look at imputing vs analysing because they're called steps somewhere, and something else somewhere else. explain what a chain is.]

1.2. Notation

The convergence guidelines introduced in this paper are developed to be integrated into the **mice** environment (Van Buuren and Groothuis-Oudshoorn 2011) in R (Core Team 2017). It therefore follows notation and conventions of Van Buuren and Groothuis-Oudshoorn (2011). For an overview of the deviations from the 'original' notation by Rubin (1987), see Van Buuren (2018).

Let Y denote an $n \times p$ matrix containing the data values on p variables for all n units in a sample. The collection of observed data values in Y is denoted as Y_{obs}; the missing part of Y is referred to as Y_{mis}. Response indicator R shows whether a data value in Y is missing or observed. The relation between R, Y_{obs}, and Y_{mis} determines the missingness mechanism. This paper only considers a 'missing completely at random' (MCAR) mechanism, where the probability to be missing is equal for all $n \times p$ cells in Y.

With multiple imputation, Y_{mis} is 'filled in' (i.e. imputed) m times. The m imputed data sets are analyzed separately, and their results pooled [Make this more clear!]. A distinction is made between the imputation step (employing p 'imputation models'), and the analysis step (using one 'complete-data model' for each of the m imputed data sets). The quantity of scientific interest is the result of the analysis step, e.g., a regression coefficient, and denoted with Q. That is, Q is estimated on each imputed data set, resulting in m separate $\hat{Q}$ values. These are combined into pooled estimate $\bar{Q}$.

2. Theoretical Background

Convergence has two interpretations [look up source!!!]: mixing between chains and stability over iterations within chains. Ideally, we want the chains to intermingle nicely, without trends within chains. We can never be certain of convergence, therefore convergence diagnostics evaluate signs of non-convergence (Hoff 2009). Diagnose non-convergence in the mixing sense: $\widehat{R}$. Diagnose non-convergence in the stability sense: auto-correlation.

In the latter interpretation of convergence, convergence is interpreted as stability over iterations, or non-recurring. Non-recurrence can be evaluated with auto-correlation. Auto-correlation shows how dependent subsequent draws of an imputation chain are on the previous value. If there is a lot of dependence, draws at e.g. iteration five are significantly correlated with the value of the first draw. Auto-correlation above the confidence level [explain that this is a critical value for correlation based on alpha .05] indicate dependence within chains. Also, possible to use MC error, Geweke, and the other one. But outside of scope of this study. Here, the focus is mainly on $\widehat{R}$.

The potential scale reduction factor $\widehat{R}$ tells us how much variance of an estimate could be shrunken down by running the chains infinitely long [Add reference, gelman something?]. That then tells us something about how dependent the chains are on the starting values. If there is no dependence on the initial values anymore, the chains have converged (in the mixing sense of the word). $\widehat{R}$ is then equal to one. Conventional threshold was $\widehat{R} < 1.2$ is acceptable. Most recent guideline by Vehtari, Gelman, Simpson, Carpenter, and Bürkner (2019) is more stringent: $\widehat{R} < 1.01$.

Why is $\widehat{R}$ not applicable on MI data? $\widehat{R}$ is not appropriate because it assumes over-dispersed initial values, which means that the initial values of the m imputation chains are 'far away' from the distribution that the chains are converging to. With MI procedures, initial values are chosen such that ... [or maybe even estimated from the observed data??]. This means that at most one initial value is necessary, but probably none at all. [I should look this up.] Without over-dispersed initial states, $\widehat{R}$ can falsely diagnose convergence (Brooks and Gelman 1998). This suggests that $\widehat{R}$ would not be sensitive enough to flag non-convergence of MI algorithms. Empirical finding, however, show that the opposite may be true: Lacerda *et al.* (2007) report $\widehat{R}$ values above the threshold of $\widehat{R} < 1.1$ after fifty iterations. Hypothesis of this simulation study is that $\widehat{R}$ will over-estimate non-convergence. [Or use:] Hypothesis based on Lacerda *et al.* (2007) is that the conventional acceptable level of $\widehat{R}$ is too strict for MI data. We expect that the simulation diagnostics will indicate valid inference before $\widehat{R}$ will.

3. Methods

This research project consists of an investigation into algorithmic convergence of MI algorithms. I will replicate Lacerda et al.'s simulation study on $\widehat{R}$ (Lacerda *et al.* 2007), and develop novel guidelines for assessing convergence. Ideally, I will integrate several diagnostics (e.g., $\widehat{R}$, *auto-correlation*, and *simulation error*) into a single summary indicator to flag non-convergence.

The aim is to evaluate whether the imputation procedure has converged. The primary research interest is in determining whether $\widehat{R}$ is an appropriate convergence diagnostic, and if so, which level of stringency suits MI data. We can apply $\widehat{R}$ to the mean (or to the first two moments) of the variables of interest. The simulation diagnostics are as recommended by Van Buuren

(2018). Namely, average bias, average confidence interval width, and empirical coverage rate (coverage probability) across simulations. Also look at distributional characteristics, and plausibility of imputed values, see Vink.

Convergence diagnostics are $\widehat{R}$ and auto-correlation (at lag one). $\widehat{R}$ is computed as follows:

> "In the equations below, N is the number of draws per chain, M is the number of chains, and $S = MN$ is the total number of draws from all chains. For each scalar summary of interest θ, we compute B and W, the between- and within-chain variances:
>
> $$B = \frac{N}{M-1}\sum_{m=1}^{M}\left(\bar{\theta}^{(\cdot m)} - \bar{\theta}^{(\cdot\cdot)}\right)^2, \text{ where } \bar{\theta}^{(\cdot m)} = \frac{1}{N}\sum_{n=1}^{N}\theta^{(nm)}, \quad \bar{\theta}^{(\cdot\cdot)} = \frac{1}{M}\sum_{m=1}^{M}\bar{\theta}^{(\cdot m)}$$
>
> $$W = \frac{1}{M}\sum_{m=1}^{M}s_j^2, \text{ where } s_m^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(\theta^{(nm)} - \bar{\theta}^{(\cdot m)}\right)^2$$
>
> ". (Vehtari *et al.* 2019, p. 5)

The proportion of within chain variance W against the weighted average of B and W, $\widehat{\text{var}}^+$ tells us how much variance of θ could be shrunken down if $N \to \infty$. $\widehat{\text{var}}^+$ is computed as follows:

$$\widehat{\text{var}}^+(\theta|y) = \frac{N-1}{N}W + \frac{1}{N}B.$$

The potential scale reduction factor $\widehat{R}$ is obtained as:

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^+(\theta|y)}{W}}.$$

[make clear when i talk about my study or lit. look at equations vehtari, because there is no real need for it. try to rephrase in my own words because of text flow.]

In this study, $\widehat{R}$ is computed for the mean of each variable, in each imputation chain, over all simulation conditions and MCMC simulations.

4. Simulation Approach

In this study, 1000 MCMC simulations are performed. The simulation set-up consists of several steps, see pseudo-code below. Simulation code is available in online appendix XYZ on Github.

```
# pseudo-code of simulation
simulate data with missingness
for (number of simulation runs from 1 to 1000)
  for (number of iterations from 1 to 100)
    impute the missingness
    compute convergence diagnostics
    perform analysis
```

```
    pool results
    compute simulation diagnostics
aggregate convergence and simulation diagnostics
```

The simulated dataset is a finite population of $N = 1000$. The data are simulated to solve a multiple linear regression complete data problem. The quantity of scientific interest is the estimated regression coefficient of predictor X on outcome variable Y. The linear regression model is $Y \sim X + Z_1 + Z_2$, where Y is the dependent variable, X is an independent variable, and Z_1 and Z_2 are covariates. The data generating model of the predictors is a multivariate normal distribution with means structure μ, and variance-covariance matrix Σ. Outcome variable Y is deduced from the predictor variables as follows:

$$\begin{pmatrix} X \\ Z_1 \\ Z_2 \\ \epsilon \end{pmatrix} \sim N \left[\begin{pmatrix} 12 \\ 3 \\ 0.5 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 4 & 1.8 & 0 \\ 4 & 16 & 4.8 & 0 \\ 1.8 & 4.8 & 9 & 0 \\ 0 & 0 & 0 & 100 \end{pmatrix} \right]$$

$$Y = 2 \times X + 0.5 \times Z_1 - 1 \times Z_2 + \epsilon$$

After data generation, the complete data is 'amputed'. That is, the **mice** function `ampute()` in R is used to impose an MCAR missingness mechanism upon the data. The missingness is univariate, and the probability to be missing is the same for all variables, namely 20%. This leaves 20% of the rows completely observed. Table XYZ shows a summary of the generated complete data, and the amputed data. The amputed dataset is the starting point of each of the 1000 simulations per simulation condition.

Missing data points are imputed with **mice** in R. All simulations are performed with imputation method 'norm' (Bayesian linear regression imputation), and five imputation chains ($m = 5$). The number of iterations is varied over simulations ('maxit' argument between 1 and 100). Each simulation condition (i.e., each number of iterations) is simulated 1000 times. Simulation and convergence diagnostics are aggregated over the 1000 MCMC simulations.

Add "post-processing steps (e.g. aggregation computations, summary statistics, regressions on the simulation results) used to transform simulation outputs to reported results".

Post-processing of simulation diagnostics:

Bias is computed as the estimated regression coefficient after MI minus the true regression coefficient. Bias is averaged over all simulations with the same simulation condition.

Confidence interval width (CIW) is computed as the diffence between the lower and upper bound of the 95% confidence interval (CI95%). The CI95% bounds are computed as the estimated regression coefficient plus or minus (respectively) the pooled SE across imputations times 2.66 (the quantile of a t distribution with m-1 degrees of freedom). CIW is averaged over all simulations with the same simulation condition.

Coverage rate is computed as the proportion of simulations (with the same simulation condition) in which the true regression coefficient is between the bounds of the CI95%.

Post-processing of convergence diagnostics:

$\widehat{R}$ is computed within imputation chains: for each variable, for each simulation condition, for each simulation. The maximum value across variables within the same simulation is reported. Maximum $\widehat{R}$ values per simulation condition are aggregated across simulations.

Autocorrelation (AC) is computed within imputation chains, as the correlation between θ at the i^{th} iteration and θ at the $i + 1^{th}$ iteration. For now, we only consider the chain mean as scalar of interest θ. The AC of the variable with the highest absolute AC value is reported per simulation. These values are then averaged per simulation condition.

5. Results

Figure 2 shows the results over 1000 simulations. A subset of simulation conditions is presented in Table 1.

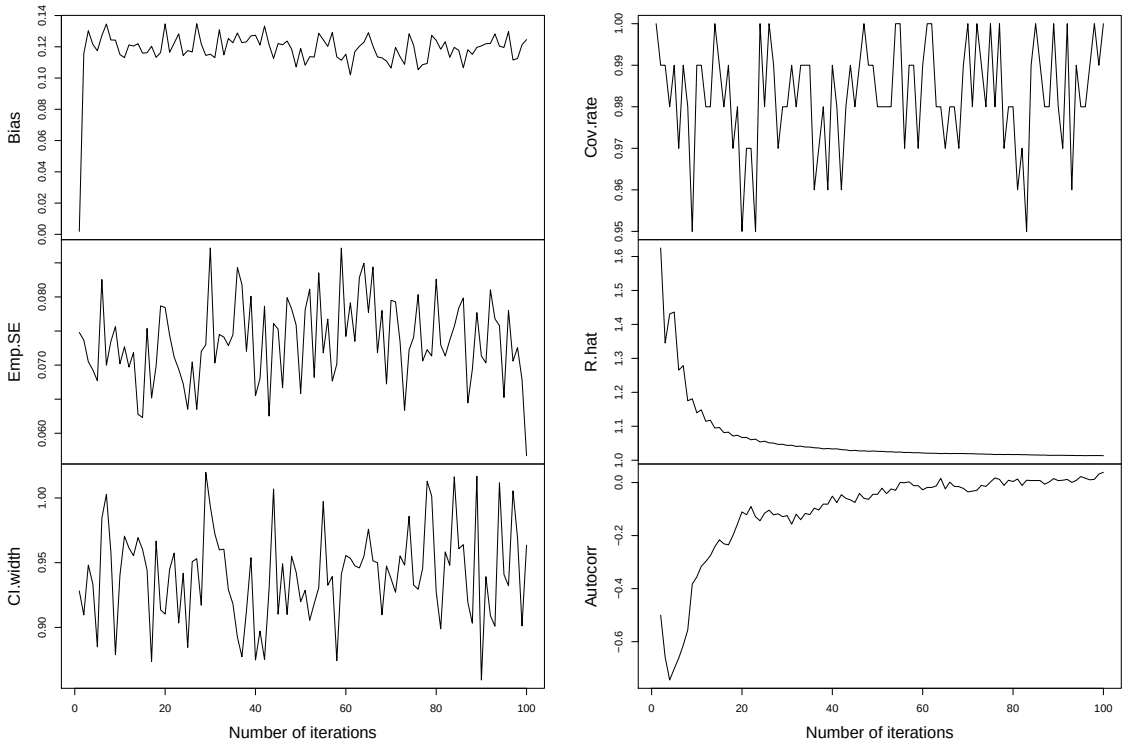

Figure 2: Simulation and convergence diagnostics over 1000 MCMC simulations.

Evaluate simulation diagnostics one by one: bias, empirical SE, confidence interval width, and coverage rate. Then evaluate convergence diagnostics $\widehat{R}$ and auto-correlation.

5.1. Simulation diagnostics

Bias. After 1 or 2 iterations bias fluctuates within a narrow range of the average bias across all simulation conditions. Average bias is not zero because there is only one incomplete dataset to be imputed. Bias is a sample effect.

Empirical standard error (SE).

Confidence interval width. Does not show a trend across the simulation conditions. [Is related to empirical SE??]

Coverage rate. Is more or less stable from two iterations upwards. But higher percentages then the expected 95% **?**. This is due to pooled SE??

5.2. Convergence diagnostics

$\widehat{R}$. Steep drop down up-to iteration 20-30, then stable. Conventional threshold $\widehat{R} < 1.2$ reached after three iterations, $\widehat{R} < 1.1$ after four iterations, $\widehat{R} < 1.01$ after 38 iterations (see online appendix).

Auto-correlation. Auto-correlation is dangerous when positive. These auto-correlations are negative. Still, we want the iterations to be stable and independent. Default maxit is five iterations now, should this be different? Are the 'waves' most pronounced at iteration 5? But why the dip??

	Simulation Diagnostics				Convergence Diagn.	
It.	Bias	Emp. SE	CI width	Cov. rate	$\widehat{R}$	Auto-corr.
1	-0.135	0.115	0.637	0.930		
2	-0.088	0.125	0.691	0.990	1.499	-0.500
3	-0.090	0.112	0.623	0.950	1.200	-0.658
4	-0.093	0.116	0.645	0.980	1.146	-0.738
5	-0.087	0.108	0.599	0.980	1.098	-0.711
6	-0.092	0.116	0.644	0.950	1.090	-0.674
7	-0.095	0.122	0.675	1.000	1.063	-0.588
8	-0.099	0.113	0.626	0.970	1.058	-0.523
9	-0.095	0.115	0.639	1.000	1.044	-0.519
10	-0.090	0.111	0.618	0.990	1.048	-0.414
15	-0.093	0.109	0.604	0.950	1.020	-0.339
20	-0.098	0.109	0.606	0.980	1.023	-0.234
25	-0.100	0.113	0.626	0.990	1.016	-0.091
30	-0.091	0.118	0.654	0.960	1.014	-0.153
40	-0.088	0.117	0.650	0.990	1.009	-0.070
50	-0.087	0.119	0.662	0.990	1.009	-0.029

Table 1: Simulation and convergence diagnostics over 1000 MCMC simulations.

From the simulation diagnostics it appears that as little as three iterations is sufficient to draw valid inference. Convergence diagnostics $\widehat{R}$ and auto-correlation, however, indicate 20-30 iterations. Apparently, they are not perfect for MI data [move interpretation to Discussion!].

[Add simulation time (computational cost) to table? Add empirical SE of all diagnostics? Sample some random simulations and check distribution and plausibility of imputations? Use chain variance as scalar of interest for rhat and autocorrelation?]

6. Summary and discussion

This note illustrates that conventional convergence diagnostics behave differently on MI data as compared to the output of 'regular' MCMC algorithms. The recommended threshold for $\widehat{R}$ is too stringent for MI data.

Moreover, $\widehat{R}$ could theoretically not be smaller than one, yet it happened several times in this study (see online appendix XYZ). How could $\widehat{R}$ smaller than 1 occur? The number of simulations is smaller than in 'regular' MCMC processes[explain why this is positive!]. Therefore, the '$(n-1/n)$' [add equation number] correction factor can influence the estimated potential scale reduction factor. This downwards bias is in the opposite direction than expected: "The mixture-of-sequences variance, V, should stabilize as a function of n. (Before convergence, we expect σ^2 to decrease with n, only increasing if the sequences explore a new area of parameter space, which would imply that the original sequences were not overdispersed for the particular scalar summary being monitored.)" (Brooks and Gelman 1998, p 438).

Future research: We would like to have convergence measures for multivariable statistics (scalars?) of interest. This is, however, dependent of the complete data model. The eigenvector decomposition method proposed by McKay (?) should be implemented. I could not find any resources to apply this method and it is outside the scope of this thesis to investigate how this approach could be implemented.

[in discussion, explain that less it is advantage of mi, not a disadv. compared to MCMC. then final par is too negative, explain scope in intro instead.]

Computational details

The results in this paper were obtained using R 3.6.1 with the **mice** 3.6.0.9000 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

This paper is written by the sole author (Hanne Oberman, BSc.), with guidance from Master thesis supervisors prof. dr. Stef van Buuren, and dr. Gerko Vink.

References

Allison PD (2001). *Missing data*, volume 136. Sage publications.

Brooks SP, Gelman A (1998). "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics*, **7**(4), 434–455. ISSN 1061-8600. doi:10.1080/10618600.1998.10474787. URL https://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474787.

Core Team (2017). *: A language and environment for statistical computing*. Vienna, Austria. Tex.organization: Foundation for Statistical Computing, URL https://www.R-project.org/.

Gelman A, Rubin DB (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science*, **7**(4), 457–472. ISSN 0883-4237, 2168-8745. doi:10.1214/ss/1177011136. URL https://projecteuclid.org/euclid.ss/1177011136.

Hoff PD (2009). *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer New York, New York, NY. ISBN 978-0-387-92299-7 978-0-387-92407-6. doi:10.1007/978-0-387-92407-6. URL http://link.springer.com/10.1007/978-0-387-92407-6.

Lacerda M, Ardington C, Leibbrandt M (2007). "Sequential regression multiple imputation for incomplete multivariate data using Markov chain Monte Carlo."

Murray JS (2018). "Multiple Imputation: A Review of Practical and Theoretical Findings." *Statistical Science*, **33**(2), 142–159. ISSN 0883-4237. doi:10.1214/18-STS644. URL https://projecteuclid.org/euclid.ss/1525313139.

Rubin DB (1987). *Multiple Imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics Applied probability and statistics. Wiley, New York, NY. ISBN 978-0-471-08705-2.

Takahashi M (2017). "Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations." *Data Science Journal*, **16**, 37. ISSN 1683-1470. doi:10.5334/dsj-2017-037. URL http://datascience.codata.org/articles/10.5334/dsj-2017-037/.

Van Buuren S (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC. ISBN 0-429-96035-2.

Van Buuren S, Groothuis-Oudshoorn K (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(1), 1–67. ISSN 1548-7660. doi:10.18637/jss.v045.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v045i03.

Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC (2019). "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC." *arXiv:1903.08008 [stat]*. ArXiv: 1903.08008, URL http://arxiv.org/abs/1903.08008.

Vink G (????). "Towards a standardized evaluation of multiple imputation routines."

White IR, Royston P, Wood AM (2011). "Multiple imputation using chained equations: Issues and guidance for practice." *Statistics in Medicine*, **30**(4), 377–399. ISSN 1097-0258. doi:10.1002/sim.4067.

Affiliation:

Hanne Ida Oberman, BSc.
Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Department of Methodology and Statistics
Faculty of Social and Behavioral Sciences
Utrecht University
Heidelberglaan 15
3500 Utrecht, The Netherlands
E-mail: h.i.oberman@uu.nl URL: https://hanneoberman.github.io/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd