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.*

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 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. 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).

Missing data is ubiquitous. MICE can solve stuff.

Validity of inference MCMC algorithms is threatened by non-convergence. But it is not known whether conventional convergence diagnostics work for MI data. Aim of this paper is to investigate convergence properties of MI algorithms by means of simulation. Results of simulation study can be used as guideline for applied researchers.

1.1. Features

The aim of this paper is to provide applied researchers a tool to assess convergence of MICE 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). All programming code used in this paper is available in the file 'XYZ.R' along with the manuscript, and on Github repository 'XYZ'.

1.2. Notation

The convergence diagnostic introduced in this paper is developed with the aim to integrate it into the R package **mice** environment. 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 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.

2. Theoretical Background

Non-convergence has two interpretations [look up source!!!]: mixing between chains and stability over iterations within chains. Diagnose non-convergence: $\widehat{R}$, auto-correlation, MC error, Geweke, the other one. However, existing methodology may not work on MI data. Focus on $\widehat{R}$.

The potential scale reduction factor tells us how much variance could be shrunken down by running the chains infinitely long. That 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).

Why is R hat not applicable on MI data? R hat is not appropriate because it assumes over-dispersed initial values, which means that the initial values of the m imputation chains should be 'far away' from the distribution that the chains are converging to (the mode of the distribution). 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 al. I should look this up. And, "if over-dispersion does not hold, σ^2_+ can be too low, which can lead to falsely diagnosing convergence." (Brooks and Gelman 1998, p 437). Empirical finding suggests otherwise: that R hat will not be smaller than 1.1 before iteration number. 50. Therefore, the aim is to replicate empirical finding Lacerda et al.

Convergence can also be interpreted as stable 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. Ideally, we want the chains to intermingle nicely, without trends within chains. Auto-correlation above the confidence level indicate dependence within chains.

3. Methods

The validity of the MI solution depends on numerous assumptions that cannot be verified from the observed data alone. So instead of statistical tests for assumptions, evaluation procedures have been developed. For the following assumptions, no reliable procedure has been proposed and/or implemented: 1) *ignorability* of the *missingness mechanism* (Rubin 1987); 2) *congeniality* of the imputation models (Meng 1994); and 3) *compatibility* of the MI modeling procedure (Rubin 1996).

The third assumption is met when the MI algorithm converges to a stable distribution. However, conventional measures to diagnose convergence– e.g., Gelman and Rubin's 1992 statistic $\widehat{R}$ –are not applicable on multiply imputed data (Lacerda, Ardington, and Leibbrandt 2007). Therefore, empirical researchers have to rely on visual inspection procedures that are theoretically equivalent to $\widehat{R}$ (White, Royston, and Wood 2011). 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).

What is already implemented?

- Trace-plots

What is not yet implemented, but exists?

- $\widehat{R}$, but too stringent (new) threshold, and assumption of over-dispersed initial values of imputation chains not met.

- Auto-correlation. Schafer (1997, p. 129) wrote on worst linear statistic. We could calculate the auto-correlation of that statistic to know that the algorithm converged elsewhere too. See autocorr function plot in SAS of worst linear function.

- Sensitivity analysis: Run algorithm several times and compare results.

What is not implemented, and NA?

- $\widehat{R}$ threshold: Replicate simulation study and build a decision rule to solve the problem with $\widehat{R}$.

- Stability of the solution: Possibly use the slope of means over iterations too to see whether there is trending. Or apply PCA on the imputed data and if that (the eigenvalues?) stays the same we know that the means and variances are stable as well, see McKay (?).

- MC error: MC error = SD/sqrt(number of iterations), where SD represents the variation across iterations. The MC error thus represents how much the means differ w.r.t. the iterations. MC error decreases as number of iterations increases. It should not be larger than 5% of the sample standard deviation.

In short, the existing literature provides both possibilities and limitations to evaluating the validity of multiply imputed data. The goal of this research project is to develop novel methodology and guidelines for evaluating MI methods, and implement these in an interactive evaluation framework for multiple imputation. This framework will aid applied researchers in drawing valid inference from incomplete datasets.

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 GR statistic $\widehat{R}$ 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 Stef van Buuren Van Buuren (2018). That is, 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.

As recommended by **?** $\widehat{R}$ is computed as .... to be able to detect ... in the tails of the distribution.

$$B = \frac{N}{M-1}\sum_{m=1}^{M}\left(\bar{\theta}^{(\cdot m)} - \bar{\theta}^{(\cdot\cdot)}\right)^2, \quad \text{where} \quad \bar{\theta}^{(\cdot m)} = \frac{1}{N}\sum_{n=1}^{N}\theta^{(nm)}, \quad \bar{\theta}^{(\cdot\cdot)} = \frac{1}{M}\sum_{m=1}^{M}\bar{\theta}^{(\cdot m)} \tag{1}$$

$$W = \frac{1}{M}\sum_{m=1}^{M}s_j^2, \quad \text{where} \quad s_m^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(\theta^{(nm)} - \bar{\theta}^{(\cdot m)}\right)^2 \tag{2}$$

$$\widehat{\text{var}}^{+}(\theta|y) = \frac{N-1}{N}W + \frac{1}{N}B \tag{3}$$

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}(\theta|y)}{W}} \tag{4}$$

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 proper imputations before $\widehat{R}$ will.

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 diagnostic
    perform analysis
    pool results
    compute simulation diagnostics
save convergence and simulation diagnostics
```

The simulated data is a finite population of $N = 1000$. The variables are dependent variable Y, independent variable X, and covariates Z_1 and Z_2. The quantity of scientific interest is the estimated regression coefficient of X on Y in linear regression model $Y \sim X + Z_1 + Z_2$. 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()` is used to impose an MCAR missingness mechanism upon the data. The probability to be missing is the same for all cells, namely XYZ%. Table XYZ shows a summary of the generated complete data, and the amputed data. The amputed data 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. The number of iterations is varied over simulations ('maxit' argument between 1 and 100). For each number of iterations, simulation and convergence diagnostics are aggregated over 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".

5. Results

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

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.

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. Therefore, the '$(n-1/n)$' correction

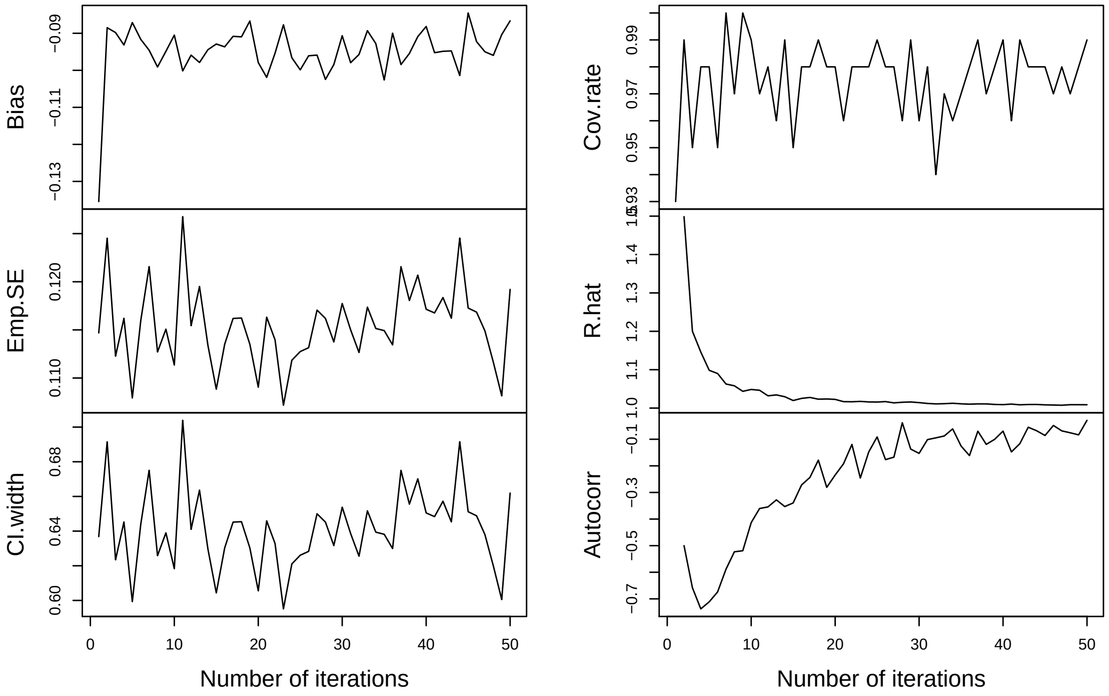

Figure 1: Simulation and convergence diagnostics over 1000 MCMC simulations.

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).

Add simulation time (computational cost) to table? Add empirical SE of diagnostics?

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.

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.

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.

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`.

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`.

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

Meng XL (1994). "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science*, **9**(4), 538–558. ISSN 0883-4237, 2168-8745. doi:10.1214/ss/1177010269. URL https://projecteuclid.org/euclid.ss/1177010269.

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.

Rubin DB (1996). "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association*, **91**(434), 473–489. ISSN 01621459. doi:10.2307/2291635. URL http://www.jstor.org/stable/2291635.

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.

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

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