Journal of Statistical Software

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

ShinyMICE: an Evaluation Suite for Multiple Imputation

Hanne Oberman
Utrecht University

Abstract

This Research Report contains the Introduction and Methods section of the technical paper that will be submitted for publication in *Journal of Statistical Software*. ("There is no page limit, nor a limit on the number of figures or tables", see https://www.jstatsoft.org/pages/view/authors.) I have chosen to use the second format from the Research Report guidelines: "It is the first half of the thesis, i.e., there are no results included yet, but the report contains a full introduction including a literature review and a methods section that contains details about the data, instruments and or statistical procedures". The goal was to develop novel methodology and guidelines for evaluating multiple imputation methods, and implement these in an interactive evaluation framework for multiple imputation: **ShinyMICE**.

Keywords: multiple imputation, evaluation methodology, **ShinyMICE**, **mice**, R.

1. Introduction: Multiple Imputation Methodology

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).

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). 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 underdeveloped in MI software, like the world leading R package **mice** (**?**). Therefore, I will answer the following question: 'Which measures are vital for evaluating the validity of multiply imputed data?'.

1.1. Features

All programming code used in this paper is available in the file XYZ.R along with the manuscript and on Github repository XYZ.

"The intended audience of this paper consists of applied researchers who want to address problems caused by missing data by multiple imputation. The text assumes basic familiarity with R. The document contains hands-on analysis using the mice package. We do not discuss problems of incomplete data in general. We refer to the excellent books by Little and Rubin (2002) and Schafer (1997). Theory and applications of multiple imputation have been developed in Rubin (1987) and Rubin (1996). van Buuren (2012) introduces multiple imputation from an applied perspective" (Van Buuren and Groothuis-Oudshoorn 2011, p. 4).

1.2. Notation and conventions?

Blue points are observed, the red points are imputed.

2. Theoretical Background

- Terminology (MCAR, MAR, MNAR)

- Rubin's rules

- FCS vs. JM?

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) *congeneality* of the imputation models (Meng 1994); and 3) *compatibility* of the MI modeling procedure (Rubin 1996).

1. A missingness mechanism is said to be ignorable when the probability to be missing does not depend on the missing data itself. Violation of this assumption can gravely affect inferences. Robustness of inferences to varying degrees of violation can be assessed with sensitivity analyses. Some practical guidelines exist (e.g., (Nguyen, Carlin, and Lee 2017)), but current MI software does not facilitate this methodology for empirical researchers.

2. Congeneal imputation models capture all required relations between observed and missing parts of the data. The extent to which this has been successful can be evaluated by plotting conditional distributions (Abayomi, Gelman, and Levy 2008). Such visualizations are available in MICE, but subsequent statistical tests to quantify the relations with covariates are not provided.

3. 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).

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.

3. Methods

Initially, the research project will consist 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.

Subsequently, I will useRShiny (Chang, Cheng, Allaire, Xie, and McPherson 2017) to implement the convergence indicator and existing evaluation measures in **ShinyMICE**, see Figure 1. The application will at least contain methodology for: sensitivity analyses; data visualizations (e.g., scatter-plots, densities, cross-tabulations); and statistical evaluation of relations between variables pre- versus post-imputation (i.e., χ^2 tests or t tests).

A working beta version of **ShinyMICE** will be considered a sufficient milestone to proceed with writing a technical paper on the methodology and the software. I will submit the paper for publication in *Journal of Statistical Software*. Finally, **ShinyMICE** will be integrated into the existing MICE environment, and a vignette for applied researchers will be written.

TheRcode and documentation of this project will be open source (available on Github). Since the study does not require the use of unpublished empirical data, I expect that the FETC will grant the label exempt.

3.1. Sensitivity Analyses

- Robustness of inferences to varying degrees of violation can be assessed with sensitivity analyses. See (Nguyen *et al.* 2017).

- Forked MICE with MNAR sensitivity analyses, see `https://thestatsgeek.com/2018/06/07/missing-not-at-random-sensitivity-analysis-with-fcs-multiple-imputation/`.

"The MAR assumption can never be tested from the observed data. One can however check whether the imputations created by MICE algorithm are plausible" (Van Buuren and Groothuis-Oudshoorn 2011, p. 42).

"Plot densities of both the observed and imputed values of all variables to see whether the imputations are reasonable. Differences in the densities between the observed and imputed values may suggest a problem that needs to be further checked" (Van Buuren and Groothuis-Oudshoorn 2011, p. 43).

3.2. Data Visualization

- Implement existing plot functions in **ShinyMICE**, add subsequent statistical tests to quantify the relations with covariates.

- Relations between observed and missing parts of the data: use **lattice** package functionalities to improve the **mice** functions `hist()` and `xyplot()`.

"An important step in multiple imputation is to assess whether imputations are plausible. Imputations should be values that could have been obtained had they not been missing. Imputations should be close to the data. Data values that are clearly impossible (e.g., negative counts, pregnant fathers) should not occur in the imputed data. Imputations should respect relations between variables, and reflect the appropriate amount of uncertainty about their 'true' values. Diagnostic checks on the imputed data provide a way to check the plausibility of the imputations" (Van Buuren and Groothuis-Oudshoorn 2011, p. 11).

"It is often useful to inspect the distributions of original and the imputed data. One way of doing this is to use the function `stripplot()` ... The differences between the red points represents our uncertainty about the true (but unknown) values. ... Under MCAR, univariate distributions of the observed and imputed data are expected to be identical. Under MAR, they can be different, both in location and spread, but their multivariate distribution is assumed to be identical." (Van Buuren and Groothuis-Oudshoorn 2011, p. 12). Create scatterplots via `xyplot()`.

3.3. Convergence Diagnostic

- Convergence issues: "Imputers who do choose to use FCS should use flexible univariate models wherever possible and take care to assess apparent convergence of the algorithm, for example by computing traces of pooled estimates or other statistics and using standard MCMC diagnostics (Gelman et al., 2013, Chapter 11). It may also be helpful to examine the results of many independent runs of the algorithm with different initializations and to use random scans over the p variables to try to identify any convergence issues and mitigate possible order dependence" (Murray 2018, p. 19).

- Replicate simulation study and build a decision rule to solve the problem: "The monitoring statistic was computed for mean monthly earnings at each iteration in chains of length k = 200. Since calculation of the statistic requires M parallel sequences, m = 5 such chains were constructed. This value of m was informed by the preferred choice given in the literature on multiple imputation. The monitoring statistic computed at each iteration is presented in Figures 15 to 17 for each of the three missingness mechanisms. The red vertical line denotes ten iterations" (Lacerda *et al.* 2007, p. 49).

- Potentially add something about updated (2019) version of $\widehat{R}$ and the new threshold of 1.01, see (Vehtari, Gelman, Simpson, Carpenter, and Bürkner 2019).

- Auto-correlation: "Applications of MICE with lowly correlated data therefore inject a lot of noise into the system. Hence, the auto-correlation over t will be low, and convergence will be rapid, and in fact immediate if all variables are independent. Thus, the incorporation of noise into the imputed data has pleasant side-effect of speeding up convergence" (Van Buuren 2018), par. 4.5. Also, schafer (1997, p. 129) wrote on worst linear statistic. We could calculate the autocorrelation of that statistic to know that the algorithm converged elsewhere too. See autocorr function plot in sas of worst linear function. Note: we're talking about missing data only, not the combined data (that autocorrelation is very high, as is the autocorrelation of deductively imputed values, like in the texp example). Worst linear function in SAS see `https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mi_sect027.htm`.

- 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. Or look at 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 general, you cannot know for sure if your chain has converged. But sometimes you can know if your chain has not converged, so we at least check for this latter possibility" (**?**, p. 101)

4. Models and software

This is a placeholder.

5. Illustrations

This is a placeholder.

6. Summary and discussion

This is a placeholder.

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 technical 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

Abayomi K, Gelman A, Levy M (2008). "Diagnostics for multivariate imputations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**(3), 273–291. ISSN 0035-9254, 1467-9876. doi:10.1111/j.1467-9876.2007.00613.x. URL http://doi.wiley.com/10.1111/j.1467-9876.2007.00613.x.

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

Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2017). "Shiny: web application framework for R." URL https://CRAN.R-project.org/package=shiny.

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.

Nguyen CD, Carlin JB, Lee KJ (2017). "Model checking in multiple imputation: an overview and case study." *Emerging Themes in Epidemiology*, **14**(1), 8. ISSN 1742-7622. doi:10.1186/s12982-017-0062-6. URL https://doi.org/10.1186/s12982-017-0062-6.

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.

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`.

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`.

A. More technical details

B. Using BibTEX

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