

Thesis Proposal

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Hanne Oberman (4216318)

03 oktober 2019

ShinyMICE: an Evaluation Suite for Multiple Imputation

Supervised by prof. dr. Stef van Buuren, & dr. Gerko Vink

Department of Methodology and Statistics

Utrecht University

Word count: 711

Introduction

At some point, any scientist conducting statistical analyses will run into a missing data problem [2]. Missingness is problematic because statistical inference cannot be performed on incomplete data, and ad hoc solutions can yield wildly invalid results [11]. To circumvent the ubiquitous problem of missing information, Rubin [9] proposed the framework of multiple imputation (MI). MI is an iterative algorithmic procedure in which missing data points are ‘guessed’ several times. The variability between these ‘guesses’ validly reflects how much uncertainty in the inference is due to missing information—that is, if all statistical assumptions are met [9].

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 [11]. 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. Therefore, I aim to answer the following question: ‘Which measures are vital for evaluating the validity of multiply imputed data?’.

The goal is to develop an MI evaluation suite for the world leading R package **MICE** [12]. This research project will aid applied researchers in drawing valid inference from incomplete data sets. Simultaneously, a contribution to the scientific literature is made by developing novel methodology and guidelines for evaluating MI data methods.

Literature Review

The numerous assumptions inherent to MI methods, can be roughly divided into assumptions about: 1) the *missingness mechanism*; 2) the relation between observed and missing parts of the data (*imputation model*); and 3) the *non-response model*. By definition, assumptions cannot be verified on missing data. Therefore, proxy measures are used.

The mechanism is assumed to be *missing at random*—i.e., the ‘cause’ of the missingness does not depend on the missing data itself [9]. Evaluation is done via sensitivity analyses to assess the robustness of inferences to violations. Some practical guidelines exist (e.g., [8]), but current MI software does not provide empirical researchers with this methodology.

The relation between observed and missing parts of the data can be evaluated by plotting conditional distributions [1]. This procedure currently lacks an interactive interface and statistical tests to quantify the relations with covariates. Moreover, model fit could be assessed using *over-imputation* [11] or *double robustness* [3].

The final assumption is met when the algorithm has converged to a stable distribution of MI values [11]. Conventional measures to diagnose convergence—e.g., Gelman and Rubin’s [5] statistic \hat{R} —do not suit multiply imputed data [6]. Therefore, empirical researchers have to rely on a visual inspection procedure that is theoretically equivalent to \hat{R} [13]. Not only is visually assessing convergence difficult for the untrained eye, it might also be futile. The convergence properties of MI algorithms lack scientific consensus [10], and some default **MICE** techniques might not converge to stable distributions at all [7]. Convergence diagnostics for MI methods have not been systematically studied [11].

Methods

This research project will be supervised by the **MICE** developers. I will develop novel methodology for evaluating MI data, and implement these methods using R Shiny [4]. The endproduct is an interactive evaluation device: ‘**ShinyMICE**’. The R code and documentation will be open source (available on Github). Since this project does not require the use of unpublished empirical data, I expect that the FETC will grant this project the label ‘exempt’.

The research report will consist of an investigation into algorithmic convergence of MI algorithms. I will replicate Lacerda et al.’s simulation study on \hat{R} [6], and potentially combine this convergence measure with

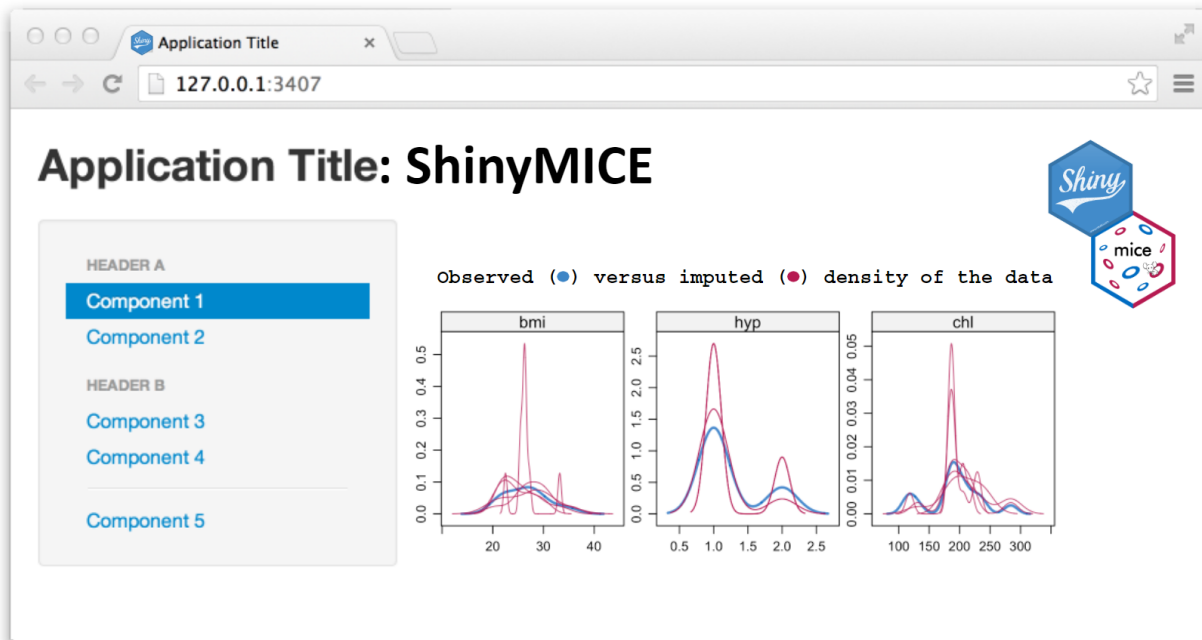


Figure 1: Preliminary impression of the interactive **ShinyMICE** user interface.

auto-correlation or simulation error diagnostics. Ideally, I will develop a single summary indicator to flag non-convergence.

Subsequently, I will implement this and other 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 evaluations 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 aim to submit this 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.

References

- [1] Abayomi, K. et al. 2008. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 57, 3 (Jun. 2008), 273–291. DOI:<https://doi.org/10.1111/j.1467-9876.2007.00613.x>.
- [2] Allison, P.D. 2001. *Missing data*. Sage publications.
- [3] Bang, H. and Robins, J.M. 2005. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*. 61, 4 (2005), 962–973. DOI:<https://doi.org/10.1111/j.1541-0420.2005.00377.x>.
- [4] Chang, W. et al. 2019. *Shiny: Web Application Framework for R*.
- [5] Gelman, A. and Rubin, D.B. 1992. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*. 7, 4 (Nov. 1992), 457–472. DOI:<https://doi.org/10.1214/ss/1177011136>.
- [6] Lacerda, M. et al. 2007. Sequential regression multiple imputation for incomplete multivariate data using Markov chain Monte Carlo. (2007).

- [7] Murray, J.S. 2018. Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*. 33, 2 (May 2018), 142–159. DOI:<https://doi.org/10.1214/18-STS644>.
- [8] Nguyen, C.D. et al. 2017. Model checking in multiple imputation: An overview and case study. *Emerging Themes in Epidemiology*. 14, 1 (Aug. 2017), 8. DOI:<https://doi.org/10.1186/s12982-017-0062-6>.
- [9] Rubin, D.B. 1987. *Multiple Imputation for nonresponse in surveys*. Wiley.
- [10] 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, (Jul. 2017), 37. DOI:<https://doi.org/10.5334/dsj-2017-037>.
- [11] Van Buuren, S. 2018. *Flexible imputation of missing data*. Chapman and Hall/CRC.
- [12] Van Buuren, S. and Groothuis-Oudshoorn, K. 2011. Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 45, 1 (Dec. 2011), 1–67. DOI:<https://doi.org/10.18637/jss.v045.i03>.
- [13] White, I.R. et al. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 30, 4 (Feb. 2011), 377–399. DOI:<https://doi.org/10.1002/sim.4067>.