Thesis Proposal

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Hanne Oberman (4216318)

2019-09-25

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: 821

Introduction

At some point, any (social) scientist conducting statistical analyses will run into a missing data problem (Allison 2001). Simply ignoring the missingness or using ad hoc solutions can yield wildly invalid inferences (Van Buuren 2018). Multiple imputation (MI; (Rubin 1987)) provides a framework to circumvent the *ubiquitous* problem of missing information. This technique – that is growing in popularity (Van Buuren 2018) – entails 'guessing' the missing values in an incomplete data set several times. The variability between the resulting completed data sets represents how much uncertainty in the inferences is due to missingness (Rubin 1987).

The R package MICE (Multiple Imputation using Chained Equations; (Van Buuren and Groothuis-Oudshoorn 2011)) is the world-leading software for multiple imputation (Van Buuren 2018). However, the package currently does not provide user-friendly means to evaluate multiply imputed data. *This is vital because of the amount of assumptions in MICE algorithms. Thanks to technological developments in computing speed and data storage it has become possible to check assumptions where that was unfeasible before. The very first assumption to check is algorithmic convergence; without convergence, all 'deeper' assumptions and subsequent inferences are invalid.*

The research question central to this project is: 'How can empirical researchers be facilitated in evaluating the validity of multiply imputed data?'. Therefore, the goal is to develop a MICE evaluation suite, featuring interactive adaptations of existing modules (like plots to compare the incomplete and completed data sets), and novel assessment tools (like a measure to flag algorithmic non-convergence). The practical relevance of this research project is in facilitating applied researchers in making valid inferences despite of missingness. Simultaneously, it contributes to the scientific literature because there is no definitive way of diagnosing non-convergence for MI algorithms.

Literature Review

Since "there is no clear-cut method for determining when the MICE algorithm has converged" (Van Buuren (2018), §6.5), users have to rely on visual inspection of the algorithm's iterations. Convergence is said to be reached when parameters (e.g., means of completed variables) are stable across iterations (White, Royston, and Wood 2011). And additionally, the variation between imputation chains should be no larger than the variation within each individual chain (Van Buuren 2018). *Conceptually, this visual evaluation procedure is similar to the 'Gelman-Rubin statistic' (Li et al. 2014). Practically, however, Gelman and Rubin's convergence diagnostic $\widehat{R}$ (1992) cannot be applied directly to MI data. The measure produces too many 'false alarms': chains that converged within ten iterations are still flagged as non-converged at iteration fifty (Lacerda, Ardington, and Leibbrandt 2007).*

Notwithstandingly, it seems like Su et al. (2011) did implement $\widehat{R}$ as convergence criterion into their R package 'mi', including the conventional cut-off. So it is worth investigating the validity of this convergence statistic in the context of multiple imputation.

The most recent literature on the topic states that convergence properties of MICE algorithms in general are still under debate (Takahashi 2017)–with specific procedures like 'predictive mean matching' posing entirely open questions (Murray 2018). Van Buuren (2018) summarizes this problem as follows: "No method works best in all circumstances. The consensus is to assess convergence with a combination of tools. The added value of using a combination of convergence diagnostics for missing data imputation has not yet been systematically studied" (Van Buuren (2018), §6.5). And that is exactly the gap in the scientific literature that this thesis will contribute to.

Methods

During this research project I will work directly with developers of the MICE R package. I will use R Shiny (Chang et al. 2019) to program an interactive evaluation device for the evaluation of multiply imputed data:

'ShinyMICE'. *This approach does not require the use of any empirical data. Therefore, the project does not need to be reviewed/approved by the faculty's ethical committee.*

The research report that is to be handed in December 13th 2019 will consist of an investigation into algorithmic convergence of MICE. Ideally, this will result in a single indicator to flag non-convergence (potentially based on $\widehat{R}$, auto-correlation, or MC error).

The second stage of the research project is of more practical nature: programming ShinyMICE. The application will at least consist of the following: 1) one or more measures to assess algorithmic convergence; 2) data visualizations (e.g., scatter-plots, densities, and cross-tabulations of the data pre- and post-imputation); and 3) statistical evaluation of relations between variables pre- versus post-imputation (i.e., χ^2-tests or t-tests). *This part of the thesis also entails research into user interface design, local versus cloud-based data storage, and optimizing the application's efficiency.*

A working beta version of ShinyMICE will be considered sufficient to proceed to the next step: writing a technical paper on the workings and usage of the software. The preferred journal for publication of this manuscript is *Journal of Statistical Software* – alternatively, publication in *The R Journal* will be attempted.

Finally, the ShinyMICE package will be integrated into the existing MICE environment, and a vignette for applied researchers will be written.

References

Allison, Paul D. 2001. *Missing Data*. Vol. 136. Sage publications.

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, Jonathan McPherson, RStudio, jQuery Foundation (jQuery library and jQuery UI library), et al. 2019. *Shiny: Web Application Framework for R* (version 1.3.2). https://CRAN.R-project.org/package=shiny.

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4): 457–72. https://doi.org/10.1214/ss/1177011136.

Lacerda, Miguel, Cally Ardington, and Murray Leibbrandt. 2007. "Sequential Regression Multiple Imputation for Incomplete Multivariate Data Using Markov Chain Monte Carlo."

Li, Fan, Michela Baccini, Fabrizia Mealli, Elizabeth R. Zell, Constantine E. Frangakis, and Donald B. Rubin. 2014. "Multiple Imputation by Ordered Monotone Blocks with Application to the Anthrax Vaccine Research Program." *Journal of Computational and Graphical Statistics* 23 (3): 877–92. https://doi.org/10.1080/10618600.2013.826583.

Murray, Jared S. 2018. "Multiple Imputation: A Review of Practical and Theoretical Findings." *Statistical Science* 33 (2): 142–59. https://doi.org/10.1214/18-STS644.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics. New York, NY: Wiley.

Su, Yu-Sung, Andrew E. Gelman, Jennifer Hill, and Masanao Yajima. 2011. "Multiple Imputation with Diagnostics (Mi) in R: Opening Windows into the Black Box" 45 (2): 1–31. https://doi.org/10.7916/D8VQ3CD3.

Takahashi, Masayoshi. 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 (0): 37. https://doi.org/10.5334/dsj-2017-037.

Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

Van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (1): 1–67. https://doi.org/10.18637/jss.v045.i03.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–99. https://doi.org/10.1002/sim.4067.