Thesis Proposal

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Hanne Oberman (4216318)

2019-09-26

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:

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).

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. Without proper evaluation of the imputations and the underlying assumptions, any drawn inference may erroneously be deemed valid (**source?**). 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 a MI evaluation suite for the world leading R package `MICE` (Van Buuren and Groothuis-Oudshoorn 2011). It will feature novel assessment tools (e.g., a measure to flag algorithmic non-convergence), and interactive adaptations of (partially) implemented modules (e.g., plots to compare the incomplete and completed data sets). 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

Elaborate: The most vital state to be evaluated is the convergence of the algorithm. Without convergence, any 'deeper' assumption and resulting inference is invalid.

Convergence properties of iterative MI algorithms are still under debate (Takahashi 2017)–with specific procedures like *predictive mean matching* posing an open question entirely (Murray 2018). Van Buuren (2018) summarizes the state of the art 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).

Currently, applied researchers have to rely on visual inspection of the algorithm's iterations (Van Buuren 2018). 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, the visual inspection procedure is equivalent to Gelman, & Rubin's (1992) convergence statistic $\widehat{R}$ (Li et al. 2014). In practice, however, $\widehat{R}$ cannot be applied directly to MI data, and is prone to producing false negatives (Lacerda, Ardington, and Leibbrandt 2007).

Add: which other assumptions could be checked? The only available paper on evaluating MI data is Abayomi, Gelman, and Levy (2008). Ideally, we would want to study all possible combinations of variables: univariate, bivariate, etc. And to include both plots and stats.

Methods

This research project will be supervised by the `MICE` developers. I will develop novel methodology for evaluating MI data, and implement these methods in an interactive evaluation device in R Shiny (Chang et al. 2019): '`ShinyMICE`'. This approach does not require the use of any empirical data. Therefore, the project is not subject to the school's ethical approval process.

The research report will consist of an investigation into algorithmic convergence of MI algorithms. Ideally, this will result in a single summary indicator to flag non-convergence (potentially based on $\widehat{R}$, auto-correlation, or MC error).

In the second stage of the research project I will focus on other evaluation measures to implement in `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 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

Abayomi, Kobi, Andrew Gelman, and Marc Levy. 2008. "Diagnostics for Multivariate Imputations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57 (3): 273–91. https://doi.org/10.1111/j.1467-9876.2007.00613.x.

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.

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.