Thesis Proposal

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Hanne Oberman (4216318)

2019-10-01

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: 775

Introduction

At some point, any (social) 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' 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.

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 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` (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

By definition, the numerous assumptions inherent to MI methods cannot be verified from the data that is missing.

- 1) assumption about the nature of the missingness (the missing data mechanism): that the 'cause' of the missingness is not missing, i.e. we can model the missingness from observed data (Rubin 1987). Robustness of the results to violations of this assumption can be assessed using sensitivity analyses. See Nguyen, Carlin, and Lee (2017) for practical guidelines.

What is missing? Implement in MICE with guidelines to interpret.

- 2) assumption about the missing data model: that we can 'guess' missing values based on the distribution of the observed values and covariates (Van Buuren 2018). This can be evaluated by evaluating plots of the distributions of the observed and imputed data. See review of diagnostics by Abayomi, Gelman, and Levy (2008).

What is missing? To develop an interactive interface to perform analyses and plot the (in)complete data, use statistical tests to assess relations between variables, and use over-imputation or 'double robustness' to evaluate model fit.

- 3) assumptions about the algorithm: that it has converged to a stable distribution (Van Buuren 2018). This is a gap in the scientific literature. There is no consensus on the convergence properties of MI algorithms (Takahashi 2017). For some MI techniques it is not known whether a stable distribution can be reached at all (Murray 2018). It is vital therefore, to assess convergence after running an MI algorithm. Multiply imputed data does not conform to the assumptions of conventional measures to diagnose convergence–e.g., Gelman and Rubin's (1992) statistic $\widehat{R}$ (Lacerda, Ardington, and Leibbrandt 2007). Empirical researchers have to rely on a visual inspection procedure that is theoretically equivalant to $\widehat{R}$ (White, Royston, and Wood 2011). In practice, however, visually assessing convergence is subjective, and difficult for the untrained eye.

What is missing? A systematic study of convergence diagnostics for MI methods is missing (Van Buuren 2018).

–>

–> –> –> –> –> –> –> –> –> –>

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 (Chang et al. 2019). 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. Ideally, this will result in a single summary indicator to flag non-convergence (potentially based on $\widehat{R}$, autocorrelation, or simulation 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).

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."

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.

Nguyen, Cattram D., John B. Carlin, and Katherine J. Lee. 2017. "Model Checking in Multiple Imputation: An Overview and Case Study." *Emerging Themes in Epidemiology* 14 (1): 8. https://doi.org/10.1186/s12982-017-0062-6.

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.