Draft Thesis Proposal

Hanne Oberman

2019-09-09

Introduction

Research goals...

develop a valid method to investigate the plausibility of multiply imputed data based on:

- models (imputation and non-response)
- data features (cross-tabs, point estimates, aggregate statistics, etc.)
- assumptions
- algorithmic convergence

Literature review

"There is no clear-cut method for determining when the MICE algorithm has converged." Van Buuren (2018), §6.5

"Convergence is diagnosed when the variance between different sequences is no larger than the variance within each individual sequence." Van Buuren (2018)

"Several expository reviews are available that assess convergence diagnostics for MCMC methods (Cowles and Carlin 1996; Brooks and Gelman 1998; El Adlouni, Favre, and Bobée 2006). Cowles and Carlin (1996) conclude that "automated convergence monitoring (as by a machine) is unsafe and should be avoided." 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)

Abayomi, Gelman, and Levy (2008)

Bartlett et al. (2015)

Li et al. (1991)

Rubin (1987)

Rubin (1996)

Vink (n.d.)

Van Buuren and Groothuis-Oudshoorn (2011)

Chang et al. (2019)

Schafer and Graham (2002)

Cowles and Carlin (1996)

Approach

R Shiny...

ShinyMICE

- single measure to assess whether algorithm converged (based on Gelman-Rubin statistic?)
- data visualizations pre and post imputation (scatterplots, densities, cross-tabs)
- statistical evaluation of relations between variables pre and post imputation (chi square or t-tests)

Initial proposal

ShinyMice: an evaluation suite for multiple imputation Supervisors: Prof.dr. Stef van Buuren, Dr. Gerko Vink

The MICE package in R is a world-leading software package for multiple imputation. When using the mice function to solve missing data, vast amounts of information are calculated and stored. However, the MICE package currently lacks a user-friendly means of assessing this information. This project focuses on developing and programming novel means of evaluating, presenting and organising this information in a web-browser based local app that allows for 1) comparing the imputed data to the observations, 2) inspecting the algorithmic convergence, 3) inspecting multivariate distributions, 4) the plausibility of the imputed data, and so on. During this project you will work closely with the developers of MICE in R and the coding components in this project will focus on R and Shiny.

The objective of the research is to 1) create a Shiny app to investigate and evaluate multiply imputed data sets, 2) implement it in MICE in R, 3) write an instructional vignette, 4) write a technical paper on the workings and usage of the software aimed at e.g. the R Journal or Journal of Statistical Software. The expected output is a state-of-the-art shiny app that can find its way into MICE.

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.

Bartlett, Jonathan W, Shaun R Seaman, Ian R White, and James R Carpenter. 2015. "Multiple Imputation of Covariates by Fully Conditional Specification: Accommodating the Substantive Model." *Statistical Methods in Medical Research* 24 (4): 462–87. https://doi.org/10.1177/0962280214521348.

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.

Cowles, Mary Kathryn, and Bradley P Carlin. 1996. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of the American Statistical Association* 91 (434): 883–904.

Li, Kim-Hung, Xiao-Li Meng, Trivellore E Raghunathan, and Donald B Rubin. 1991. "Significance Levels from Repeated P-Values with Multiply-Imputed Data." *Statistica Sinica*, 65–92.

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.

———. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91 (434): 473–89. https://doi.org/10.2307/2291635.

Schafer, Joseph L, and John W Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147.

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.

Vink, Gerko. n.d. "Towards a Standardized Evaluation of Multiple Imputation Routines."