Research Report

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Hanne Oberman (4216318)

17 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

Requirements

Goals:

- find an answer to a research question in the field of Methods and Statistics
- get work experience in the field of Methods and Statistics
- a first step in writing a scientific paper: perform research and report the findings in a short research report

Grading:

- The amount of work done
- Work independency
- Clarity and logic of the research report (clear and correct formulations, a logical structure of sections and of paragraphs within sections, no typos or errors in grammar, and flow of sections, like key questions from literature or conclusions from results).
- Nice looking and consistent lay-out of the report (conciseness in presentation of tables, figures, and formulas (if any)).
- General comments

Format (either/or):

- It is written as a (mini) thesis, with an introduction, methods section, some results (i.e., preliminary analyses, or pilot simulations), and a discussion of results. The length of the research report should be maximally 2500 words of text (without references list and/or tables and figures). Please do not include appendices, and no more than 6 tables/figures. Table and Figure captions do not count towards the word limit. An abstract may be included, but is not necessary.
- It is the first half of the thesis, i.e., there are no results included yet, but the report contains a full introduction including a literature review and a methods section that contains details about the data, instruments and/or statistical procedures. There is no (absolute) word limit but the length of the report should comply with recommendations/expectations of the proposed journal. Note that most journals have word limits or rather strong opinions on (too) lengthy manuscripts.

Introduction

- MI (and ad hoc solutions?)

From thesis proposal:

At some point, any 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 (Rubin 1987) proposed the framework of multiple imputation (MI). MI is an iterative algorithmic procedure in which missing data points are 'guessed' (i.e. imputed) several times. The variability between the imputations validly reflects how much uncertainty in the inference is due to missing information–that is, if all statistical assumptions are met (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* (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, like the world leading `R` package `MICE` (Van Buuren and Groothuis-Oudshoorn 2011). Therefore, I will answer the following question: 'Which measures are vital for evaluating the validity of multiply imputed data?'.

Theoretical Background

- Terminology (MCAR, MAR, MNAR)
- Rubin's rules
- FCS vs. JM

From thesis proposal:

The validity of the MI solution depends on numerous assumptions that cannot be verified from the observed data alone. So instead of statistical tests for assumptions, evaluation procedures have been developed. For the following assumptions, no reliable procedure has been proposed and/or implemented: 1) *ignorability* of the *missingness mechanism* (Rubin 1987); 2) *congeneality* of the imputation models (Meng 1994); and 3) *compatibility* of the MI modeling procedure (Rubin 1996).

1. A missingness mechanism is said to be ignorable when the probability to be missing does not depend on the missing data itself. Violation of this assumption can gravely affect inferences. Robustness of inferences to varying degrees of violation can be assessed with sensitivity analyses. Some practical guidelines exist (e.g., (Nguyen, Carlin, and Lee 2017)), but current MI software does not facilitate this methodology for empirical researchers.
2. Congeneal imputation models capture all required relations between observed and missing parts of the data. The extent to which this has been successful can be evaluated by plotting conditional distributions (Abayomi, Gelman, and Levy 2008). Such visualizations are available in `MICE`, but subsequent statistical tests to quantify the relations with covariates are not provided.[1]
3. The third assumption is met when the MI algorithm converges to a stable distribution. However, conventional measures to diagnose convergence– e.g., Gelman and Rubin's (1992) statistic $\widehat{R}$ –are not applicable on multiply imputed data (Lacerda, Ardington, and Leibbrandt 2007). Therefore, empirical researchers have to rely on visual inspection procedures that are theoretically equivalent to $\widehat{R}$ (White, Royston, and Wood 2011). Visually assessing convergence is not only difficult to the untrained eye, it might also be futile. The convergence properties of MI algorithms lack scientific consensus (Takahashi 2017), and some default `MICE` techniques might not converge to stable distributions at all (Murray

[1] Additionally, there is potential to assess model fit by means of *over-imputation* (Van Buuren 2018) or *double robustness* (Bang and Robins 2005)– topics that are only persued if time permits.

2018). Moreover, convergence diagnostics for MI methods have not been systematically studied (Van Buuren 2018).

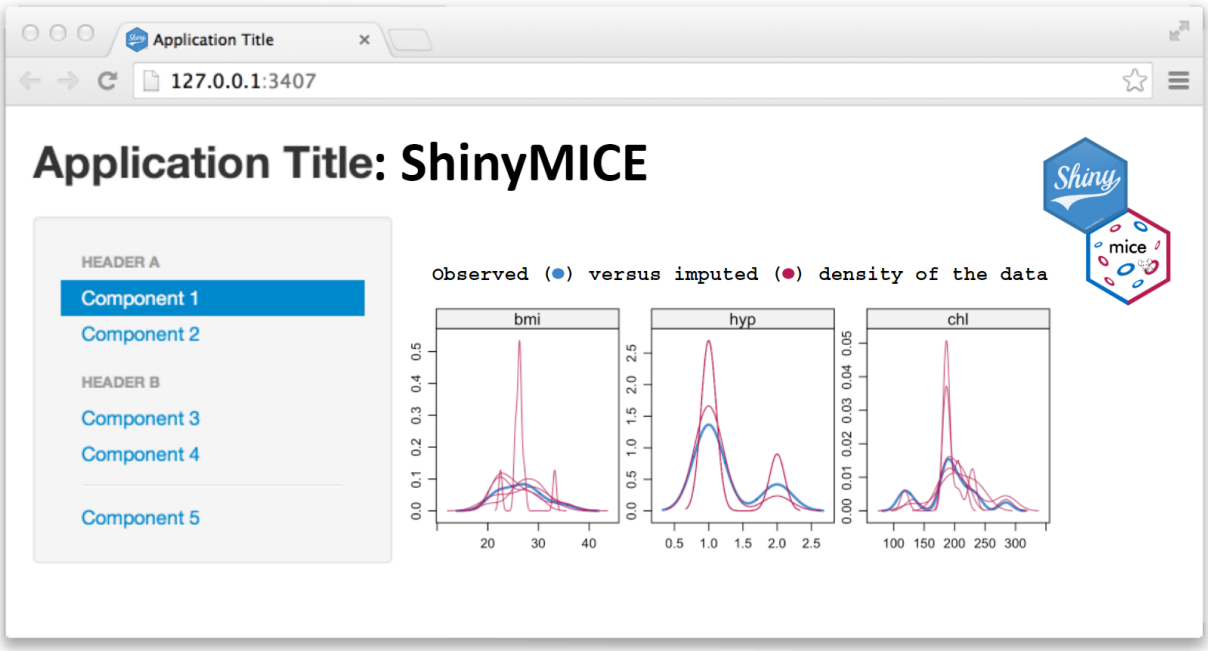
In short, the existing literature provides both possibilities and limitations to evaluating the validity of multiply imputed data. The goal of this research project is to develop novel methodology and guidelines for evaluating MI methods, and implement these in an interactive evaluation framework for multiple imputation. This framework will aid applied researchers in drawing valid inference from incomplete datasets.

Methods

From thesis proposal:

Initially, the research project will consist of an investigation into algorithmic convergence of MI algorithms. I will replicate Lacerda et al.'s simulation study on $\widehat{R}$ (Lacerda, Ardington, and Leibbrandt 2007), and develop novel guidelines for assessing convergence. Ideally, I will integrate several diagnostics (e.g., $\widehat{R}$, *auto-correlation*, and *simulation error*) into a single summary indicator to flag non-convergence.

Subsequently, I will use `R Shiny` (Chang et al. 2019) to implement the convergence indicator and existing 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 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 will submit the paper 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.

The R code and documentation of this project will be open source (available on Github). Since the study does not require the use of unpublished empirical data, I expect that the FETC will grant the label 'exempt'.

Replicate Lacerda, Ardington, and Leibbrandt (2007): chains that converged within ten iterations are still flagged as non-converged at iteration fifty (Lacerda, Ardington, and Leibbrandt 2007).

"The monitoring statistic was computed for mean monthly earnings at each iteration in chains of length k =

200. Since calculation of the statistic requires M parallel sequences, m = 5 such chains were constructed. This value of m was informed by the preferred choice given in the literature on multiple imputation. The monitoring statistic computed at each iteration is presented in Figures 15 to 17 for each of the three missingness mechanisms. The red vertical line denotes ten iterations." (Lacerda, Ardington, and Leibbrandt 2007, 49)

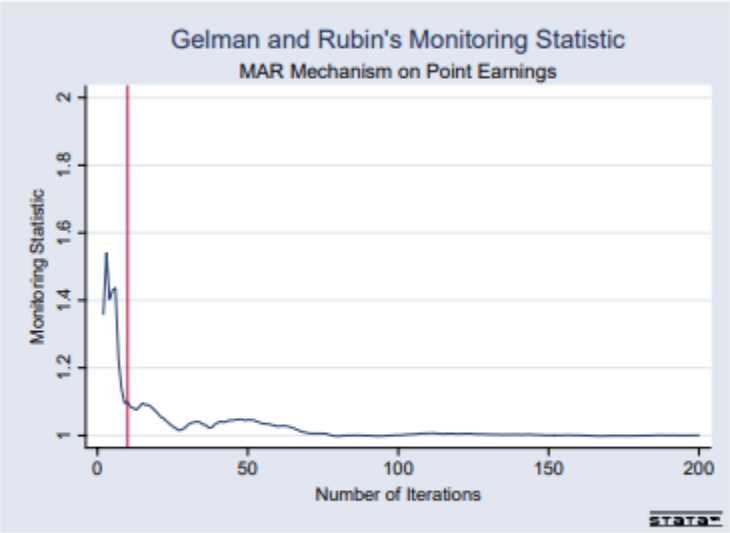


Figure 16: Gelman and Rubin's monitoring statistic at each iteration under an MAR mechanism. Red vertical line corresponds to 10 iterations.

- Notes MICE meetings:

Why do people say the GR stat will give false positives? perhaps the window (nr of it) is too small. if you start from dispersed, you'll end up with convergence (GR = 1) –> the bad thing is to stop too early. this is a false negative, not positive! possible approach: replicate and build a decision rule to solve the problem.

In mice algorithms, iterating is very fast compared to regular MCMC. therefore, we could use the average over the last 5 iterations you'll take the 'majority vote' to indicate conv. so compute GR at each iteration: band and threshold. another approach is to compute autocorrelation (which is decreasing over itarations because we add in noise. schafer (1997, p. 129) wrote on worst linear stat and calculate the autocorr of that stat we know that the alg did converge. mice has noise so little autocorr. but we have very rich data so we're not estimating everything from scratch. not: we're talking about missing data only, not the combined data (that autocorr. is very high). autocorr is high if we can predict the miss value very well. the higher r sq the higher the autocorr. so the harder to converge. watch out for the problem with texp that will converge at once because you can deductively impute the value per class. possibly use slope of mean too to see whether there is trending. maybe also compute first eigenvalue of the matrix, not just mean and sd. also in pca we have several measures for nr of components, why not do something similar for mice? however, now we use mean and sd of missing values only, and that's not possible fr correlations because we'll have different ns for different correls. but of we take the compl. data that should not be a problem because that's constant. maybe look into allpying pca on the imputed data and if that stays the same we know that the means and variances are stable too. patrick ... and ian white work on stata proc mi also work on convergence diagnostics –> autocorr function plot in sas of worst linear function, and likelihood ratio that you could save.

- Bayes course on convergence:

```
# E. Assessing convergence
# ------------------------------------------------------------------------------
# History plot, autocorrelation plot
# History plots show the sampled parameters over the iterations (excluding
# the burn-in).
# The development/pattern in these plots gives an indication of convergence.
# When the history plot is stable (a fat catterpillar), convergence is reached.
# Autocorrelation can be measured at many lags.
# High autocorrelation indicates slow mixing of the random path.

# mcmcplot(mcmcout = samples)
# #sigma seems fine, but b doesn't:
# #even at lag 5 there is still quite some autocorrelation: therefore we need to center the predictors!
# #otherwise we could use a million iterations to deminish this effect

# Gelman-Rubin diagnostic
# The Gelman and Rubin statistic requires you run
# the sampler/algorithm at least twice: These runs are
# referred to as multiple chains.
# It compares the variance between chains to the variance
# within chains (G-R statistic = T/W = (pooled within chain
# var + between chain var)/pooled within chain var.
# It's the red line in the plot, and should be near 1.
# See Gibbs sampler presentation (week 2)
# https://drive.google.com/file/d/1ABHm8ala3c_puVvf32zF8h6TOZ3898uB/view
# pdf p. 41, slide nr 29.
# gelman.plot(samples)
# #this appears to be okay with a really small deviation from 1

# MC error (OPTIONAL?)
# MC error = SD/sqrt(number of iterations)
# SD represents the variation across iterations
# MC error thus represents how much the means differ w.r.t. the iterations
# MC error decreases as number of iterations increases.
# It should not be larger than 5% of the sample standard deviation

# History and density plot (OPTIONAL)
# plot(samples)

# Autocorrelation plots (OPTIONAL)
# autocorr.plot(samples)

# If parameters did not converge, you may:
# . Use (many) more iterations
# . Use a different parametrization (e.g., center predictors)
# . Use different priors (e.g., multivariate normal prior (i.e.,
#   dmnorm(,)) for parameters which are correlated)
# . Use other initial values
```

Additional resources (not used)

quick win: xyplot and hist have all lattice functions (all sorts of scatterplots!)

Potentially add something about updated (2019) version of $\widehat{R}$ and the new threshold of 1.01, see Vehtari et al. (2019). Or leave it for the Research Report.

"For models based on multiple imputed data sets, this [R hat > 1.1] is often a false positive: Chains of different submodels may not overlay each other exactly, since there were fitted to different data" (see https://cran.r-project.org/web/packages/brms/vignettes/brms_missings.html).

"Imputers who do choose to use FCS should use flexible univariate models wherever possible and take care to assess apparent convergence of the algorithm, for example by computing traces of pooled estimates or other statistics and using standard MCMC diagnostics (Gelman et al., 2013, Chapter 11). It may also be helpful to examine the results of many independent runs of the algorithm with different initializations and to use random scans over the p variables to try to identify any convergence issues and mitigate possible order dependence" (Murray (2018), p. 19).

Gelman and Rubin (1992)'s convergence criterion is defined as the ratio between the sum of the pooled within chain variance plus the between chain variance, and the pooled within chain variance.

"Why can the number of iterations in MICE be so low? First of all, realize that the imputed data Ymis form the only memory in the the MICE algorithm. Chapter 3 explained that imputed data can have a considerable amount of random noise, depending on the strength of the relations between the variables. Applications of MICE with lowly correlated data therefore inject a lot of noise into the system. Hence, the auto-correlation over t will be low, and convergence will be rapid, and in fact immediate if all variables are independent. Thus, the incorporation of noise into the imputed data has pleasant side-effect of speeding up convergence" (Van Buuren (2018), par. 4.5).

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.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61 (4): 962–73. https://doi.org/10.1111/j.1541-0420.2005.00377.x.

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."

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9 (4): 538–58. https://doi.org/10.1214/ss/1177010269.

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.

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

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 (July): 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.

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2019. "Rank-Normalization, Folding, and Localization: An Improved $\widehat{}R{}$ for Assessing Convergence of MCMC," March. http://arxiv.org/abs/1903.08008.

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.