Missing the Point: Convergence of Multiple Imputation using Chained Equations Algorithms

Hanne Oberman

16-3-2020

Introduction

Feedback:

- Focus on convergence, not evaluation of MI in general.
- Combine Theoretical Background and Intro.

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 without employing *ad hoc* solutions (e.g., list-wise deletion), which may yield wildly invalid results (Van Buuren 2018). A popular answer to the ubiquitous problem of missing information is to use the framework of multiple imputation (MI), proposed by Rubin (1987). MI is an iterative algorithmic procedure in which missing datapoints are 'imputed' (i.e. filled in) several times. The variability between imputations is used to reflect how much uncertainty in the inference is introduced by the missingness. Therefore, MI can provide valid inferences despite missing information (**maybe add that this refers to CIs/p-values**).

To obtain valid inferences with MI, the variability between imputations should be properly represented (Rubin 1987; Van Buuren 2018). If this variability is under-estimated, confidence intervals around estimates will be too narrow, which can yield spurious results. Over-estimation of the variance between imputations results in unnecessarily wide confidence intervals, which can be costly because it lowers the statistical power (**maybe add why this is costly**). Since both of these situations are undesirable, imputations and their variability should be evaluated. Evaluation measures, however, are currently missing or under-developed in MI software, like the world-leading `mice` package (Van Buuren and Groothuis-Oudshoorn 2011) in `R` (R Core Team 2019). The goal of this research project is to develop novel methodology and guidelines for evaluating MI methods. These tools will subsequently be implemented in an interactive evaluation framework for multiple imputation, which will aid applied researchers in drawing valid inference from incomplete datasets.

This note provides the theoretical foundation towards the diagnostic evaluation of multiple imputation algorithms. For reasons of brevity, we only focus on the MI algorithm implemented in `mice` (Van Buuren and Groothuis-Oudshoorn 2011). The convergence properties of this MI algorithm are investigated through model-based simulation.[1] The results of this simulation study are guidelines for assessing convergence of MI algorithms.

Terminology

This note follows notation and conventions of `mice` (Van Buuren and Groothuis-Oudshoorn 2011). Basic familiarity with MI methodology is assumed. For the theoretical foundation of MI, see Rubin (1987). For an accessible and comprehensive introduction to MI from an applied perspective, see e.g. Van Buuren (2018).

[1] All programming code used in this note is available from [github.com/gerkovink/shinyMice/simulation]{https://github.com/gerkovink/ShinyMICE/simulation}.

Let Y denote an $n \times p$ matrix containing the data values on p variables for all n units in a sample. The collection of observed data values in Y is denoted by Y_{obs}, and will be referred to as 'incomplete' or 'observed' data. The missing part of Y is denoted by Y_{mis}. Which parts of Y are missing is determined by the 'missingness mechanism'.
This note only considers a 'missing completely at random' (MCAR) mechanism, where the probability of being missing is equal for all $n \times p$ cells in Y (Rubin 1987).

Figure 1 provides an overview of the steps involved with MI—from incomplete data, to m completed datasets, to m estimated quantities of interest ($\hat{Q}$s), to a single pooled estimate $\bar{Q}$. This note focuses on the algorithmic properties of the imputation step.

The MI algorithm in `mice` has an iterative nature. For each missing datapoint in Y_{mis}, m 'chains' of potential values are sampled. Only the ultimate sample that each chain lands on is imputed. The number of iterations per chain will be denoted with T, where t varies over the integers $1, 2, \ldots, T$. This is repeated m times. Imputed values are the ultimate samples of a 'chain' of For each imputed value ($t = T$), a s for each missing datapoint in Y_{mis}, . Each of the m chains starts with an initial value, drawn randomly from Y_{obs}. The chains are terminated after a predefined number of iterations. , and subsequently used in the analysis and pooling steps. The collection of samples between the initial value (at $t = 1$) and the imputed value (at $t = T$) will be referred to as an 'imputation chain'.

Methods

Results

Feedback:

- Remove the table.
- Add more info about figure legends and axes.

Discussion

References

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

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. https://www.R-project.org/.

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.

Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. Chapman; 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.