Missing Data Analysis
Advanced Introductory Reading List

Compiled by Kyle M. Lang

2019-10-02

Seminal Books

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 519). New York, NY: John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data* (Vol. 72). Boca Raton, FL: Chapman & Hall/CRC.

Generalist Books

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage Publictions.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: The Guilford Press.

This is a very accessible book that does also cover technical details of missing data handling strategies in an approachable way, without dumbing them down.

- **Missing data mechanisms** - The explanation of the mechanisms is a very approachable and non-technical. However, the book provides a nice overview of how to **test for the MCAR** (see section 1.9).
- **Traditional imputation methods** - Chapter 2 is entirely dedicated to describing these methods in great detail, and it provides great insights in their problems (e.g. biasing estimates, reduction of SE).
- **Maximum Likelihood Missing Data Handling** - The book dedicates chapter 4 to the detailed description of this handling technique, the only other acceptable alternative to multiple imputation (see also chapter 3 for a good overview of Maximum Likelihood estimation in general).
- **Multiple Imputation** - The discussion of multiple imputation is mainly divided into two chapters, 7 and 8, where the author describes the imputation, and the analysis and pooling phases, respectively, in detail. The imputation phase is described according to the **data augmentation**

algorithm [1]. Great attention is payed to the Bayesian nature of this algorithm (there even is an introductory chapter on Bayesian statistics). The pooling phase describes both the pooling of point estimates (section 8.3), and standard errors (section 8.5), along with (multi)parameter significance testing (D1, D2, etc. [see sections 8.11 through 8.13].

- **Diagnostic techniques** - Great attention is dedicated in chapter 7 to the description of convergence diagnostics in the imputation phase. In particular, the meaning of convergence in this context [section 7.7], time-series plots [section 7.9], worst linear function [section 7.9], and auto-correlation function plots (correlogram) [section 7.10], (exploratory) multiple chains [section 7.11] are all discussed in details and w/ examples. In section 9.2 there is also a detailed description of some solutions to convergence issues (and related *too-many-columns* issues), like the **ridge prior** for the covariance matrix, a semi-informative prior that adds some information coming from imaginary data records to reduce the impact of too many variables.
- **Dimensionality issues** (p>n) Apart from the description of the ridge prior for covariance matrix in section 9.2, the book also describes specific methods for questionnaires imputations (i.e. scale-level and duplicate-scale imputation, and three-step item level imputation [section 9.6]

Graham, J. W. (2012). *Missing data: Analysis and design.* New York, NY: Springer. doi: 10.1007/978-1-4614-4018-5

MI-Focused Books

Carpenter, J., & Kenward, M. (2013). *Multiple imputation and its application.* Chichester, West Sussex: John Wiley & Sons.

van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, FL: CRC Press. doi: 10.1201/b11826

This book is practically a technical companion to the R package that implements the MICE algorithm.

- The book describes the **missing data mechanisms** following the classification of **?** and describing the missing data models in accessible probabilistic terms (see sections 1.2 and 2.2). The concept of ignorability is also precisely defined in section 2.2.5 and 2.2.6 [see also 6.2].
  A guide to **generating MAR** data is presented in section 3.2.4 (bivariate) and 3.2.5 (multivariate missingness).
- **Missing data handling techniques** - The bulk of the book is focused on multiple imputation methods (maximum Likelihood is not described, and

[1] There is a some confusion on how the use of this term to refer to the this imputation algorithm. In particular it did not seem to be consistent w/ how **?** used it. Pay attention to this in future readings

very little space is dedicated to traditional methods on ad hoc solutions [see section 1.3]), and in particular to the MICE algorithm.
The treatment of multiple imputation is dealt with in first in the case of univariate missingness (ch. 3) and then extended to a multivariate scenario (ch. 4).
  - Chapter 3 - Focuses on the ability of MI to include not only noise around the prediction but also model (parameters) uncertainty when using normal models. Bayesian multiple imputation and bootstrap multiple imputation are presented as alternative ways of implementing the "**predict + noise + parameters uncertainty**" method [see section 3.2.2]. Other approaches presented here: predictive mean matching, and classification and regression trees.
  - Chapter 4 - Focuses on multiple imputation techniques that deal with multivariate missing data. The main difference w/ respect to univariate cases is that multiple patterns of missingness are possible when more variables are missing a value. The chapter describes first the types of missing data patterns, and describes how to use MICE to detect/describe these patterns. The rest of the chapter is dedicated to describing three approaches to multivariate missing data patterns: monotone data imputation, joint modelling (**JM**), and full conditional specification (**FCS**). FCS is the preferred approach by the book. The MICE algorithm is one way of implementing the FCS method.
- Analysis and **pooling phase** - These are dealt with in chapter 5. In particular, section 5.3 deals with multiparameter inference describing D1 (multivariate Wald test), D2 (combining test statistics, in general not specifically Wald test), and D3 (likelihood ratio tests).
- **Diagnostic techniques** are also well discussed (Ch. 6) - Two particular aspects are discussed:
  - Algorithmic convergence diagnostic (section 6.5.2)
  - Model fit diagnostic (section 6.6), done according to the concept of *distributional discrepancy* between imputed and observed data and performed through the use of diagnostic plots.
- **Dimensionality issues** (p>n) - The text deals with the issue of dimensionality mainly by guiding the researcher through the process of reducing the number of variables to be included in the model rather than suggesting any way of dealing with the issue at hand (see for example section 9.1 and 11.2)

Seminal Papers

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*(278), 200–203.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38. doi: 10.2307/2984875

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the sixth berkeley symposium on mathematical statistics and probability, volume 1: Theory of statistics*.

The authros provide a technical description of the Missing Information Principle defined through the decompostion of the infomration matrix for a vector of parameters of interest θ. The lost (missing) information is defined as the difference between the information matrix for θ in the hypotetical complete dataset, and the information matrix for θ obtained with just the observed cases. This decomposition is also used to describe the increase in variance in the parameter estimates caused by the missing data.

For a more approachable definition of the Missing Information Principle see Savalei and Rhemtulla (2012).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys – A phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association* (Vol. 1, pp. 20–34).

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 477–494.

This article provides a short yet insightgul review of the Missing Information Principle (an interested reader may want to consult Orchard and Woodbury (1972) for a more technical presentation of the same concept). According to it, the missing information for a parameter estimation due to missing cases is equal to the difference between the complete-data information matrix and the observed-data information matrix.

Subsequently, the definition of the Fraction of Missing Information (FMI) is exploared under both the Maximum Likelihood and the Multiple Imputation framework. The fundamental contribution of this paper is showing how FMI is not a prerogative of MI.

A final contribution is a detailed discussion of three different possible interpretations of FMI: (relative) loss of (estimation) efficacy, loss of statistical power; width inflation factor (i.e. how much bigger are the confidence intervals).

Important Algorithms

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561–581. doi: 10.1111/j.1540-5907.2010.00447.x

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

In this paper, Raghunathan *et al* introduce the SRMI (Sequential Regression Multivariate Imputation). The apporach imputes missing values on a variable-by-variable basis, by using posterior predictive disitrbutions of the missing data, conditional on the observed data. Apart from descirbing the principles, strengths, and weakness of the appraoch, their work also provides detailed instructions on how to perform the imputation according to SRMI (see Appendix A for a precise discussion on how to draw from a variety of regression models supported by SRMI). Two case-studies and one simulation study are contextually presented and they are instrumental in showing how the SMRI apporach works compared to complete-case analysis and the Multiple Imputation based on a joint multivariate model.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivarite imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

Algorithms to perform Multiple Imputation by FCS with different types of dependent vairables are presented in Appendix A of this paper. Appendix B describes a method to generate nonmonotone multivariate missing data under MAR. In the main text, after a concise introduction to imputation by FCS, a brief but rigorous definition of *compatibility* between conditonal distributions is given in this paper. A good technical complement for an interested read is **?**. Simulations and study cases are then discussed in detail for both univariate and multivariate cases, accounting for different variable types (i.e. continous, dichotomous, and polytomous). These studies highlight the performances of FCS multiple imputation as compared to complete-case analysis in terms of bias and coverage of the confidence intervals. Furthermore, a simulation study to assess the consequences of (in)compatibility. This is just an exemplifying scenario: not all possibile incompatibility schemes can be considered. Yet, it does provide compelling evidence in favor of FCS being robust to (clear) incompatibility.

Reviews/Tutorials

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40–64.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

This work reviews the state of multiple imputation 18 years after the publication of his seminal works published in 1976 and 1978. This paper clarifies the two main goals that Multiple Imputation was designed to achive: the basic objective of allowing the data ultimate users to apply the same analytical methods as if the data had been complete; and the supplemental objective of granting analyses that are statistically valid for a scientifc estimand'. The concept of statistcal validity is operationalised with clarity by distinguishing between two specification: randomization validity and confidence validity. Furthermore, the concept of *proper* mulitple imputation is exapanded upon by summarizing its main requirements. Finally some criticisms to MI are addressed by using these concepts.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, *338*, b2393.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399.

White's contribution provides some useful comparative guidelines and recommendations on different aspects of MI:

- Handling different dependent variable distirbutions/natures. In particular, there is a helpful section dealing with how to handle imputation for skewed continuous vairables.
- Variable selection - What to include and, more importantly hot to include all the variables of the analysis model in the imputation model (i.e. allowing for qudratic terms and interactions). Three main ways of doing so are presented (i.e. passive apporach, improved passive apporach using PMM, and JAV). Altought this section does provide a good comparative overview of the methods, it is not particularly helpful in describing the methods for a novice audience. Better resources that do accomplish such task are: INSERT REF HERE.
- Number of imputations – The authors criticize the *efficiency argument* (that usually leads to the rule of thumb $m = 5$ is adequate for FMI $\leq .25$) thorguh the *replicability argument*, centering the discussion around the decision of m in terms of Monte Carlo errors. The authros's rule of thumb is that $m \geq$ of incomplete cases.
- Limitations and pitfalls of MI – Apart from the lack of theoretical background, the authors discuss the following pitfalls: perfect prediction (potentially a problem when the dependent variable is categorical); sensitivity to MAR violation, non-convergence issues; and the problem of too many variables.

Important Developments/Clarifications/Extensions

Collins, L. M., Schafer, J. L., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Pscyhological Methods*, *6*(4), 330–351. doi: 10.1037//1082-989X.6.4.330

Collins *et al* investigate the effects of an inclusive *vs* a restrictive strategy to the use of auxiliary vairables in the missing data handling model of either a ML (maximum likelihood) or MI (Multiple Imputation) approach.

The authors distinguish between three categories of auxiliary vairables (depending on their correlation with the variable(s) with missing values of interest and the missingness mechanism).

Thorugh 4 different simulation setups, the they provide insights on the consequences of including or excluding auxiliabry variables, from each of these categories, and ultimetly provide compelling evidence to prefer an inclusive strategy.

A final point that is stressed by the artcile is that these effects heavely depend on the type of MAR. For example, in their simulations they found that parameter estiamte of the mean value a variable Y of interest is biased when the auxiliary vairable Z is excluded from the data handling model if the probability of missingess is linearly related to Z (MAR-linear); however, that's not the case when the probability of missigness is larger for extreme values of Z or is a function of the correlation between Z and Y.

An interesting point is raised regarding the ease of implementation of the inclusive strategy. In particular, the authors point out that, while the inclusion of auxiliary variables is straightforward in the MI framework, current (up to 2001) softwere implementations of ML methods do no facilitate it. This topic is thoroughly explored by Graham (2003).

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 80–100. doi: 10.1207/S15328007SEM1001_4

The author effectively showed thorugh multiple simulation studies that including auxiliary vairables, when using a FIML apporach to handling and analysing datasets with missing data points, can be done relativly easly under a structural equation modelling framework.

With this paper, Graham introduced the *Saturated Correlated model* and *the extra dv model* to include auxiliary vairbales in a SEM model. Ultimately, he showed that it is possible to include auxiliary vairables in a ML missing data handling procedure, without affecting the substantive model (that is to say obtaining the same parameter estimates, standard errors, and estimates of quality of fit).

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the sixth berkeley symposium on mathematical statistics and probability, volume 1: Theory of statistics.*

The authros provide a technical description of the Missing Information Principle defined through the decompostion of the infomration matrix for a vector of parameters of interest θ. The lost (missing) information is defined as the difference between the information matrix for θ in the hypotetical complete dataset, and the information matrix for θ obtained with just the observed cases. This decomposition is also used to describe the increase in variance in the parameter estimates caused by the missing data.

For a more approachable definition of the Missing Information Principle see Savalei and Rhemtulla (2012).

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 477–494.

This article provides a short yet insightgul review of the Missing Information Principle (an interested reader may want to consult Orchard and Woodbury (1972) for a more technical presentation of the same concept). According to it, the missing information for a parameter estimation due to missing cases is equal to the difference between the complete-data information matrix and the observed-data information matrix.

Subsequently, the definition of the Fraction of Missing Information (FMI) is exploared under both the Maximum Likelihood and the Multiple Imputation framework. The fundamental contribution of this paper is showing how FMI is not a prerogative of MI.

A final contribution is a detailed discussion of three different possible interpretations of FMI: (relative) loss of (estimation) efficacy, loss of statistical power; width inflation factor (i.e. how much bigger are the confidence intervals).

Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*(1), 265–291.