Missing Data Analysis
Advanced Introductory Reading List

Compiled by Kyle M. Lang & Edoardo Costantini

2020-03-02

The sources listed below offer an extended/advanced introduction to missing data analysis. This list is meant to cover the basics of missing data from a technical, mathematically rigorous perspective. Readers without a strong background in mathematics/statistics may wish to begin with the sources listed in "gentle_intro.tex".

Most of the following sources represent either seminal references in missing data or general overviews of the field. Consequently, this list does not necessarily represent the latest work in missing data theory. These sources should, however, provide a very thorough introduction to/overview of modern missing data theory.

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, to some extent, does cover technical details of missing data handling strategies in an approachable way. The following topics are explored by this resource:

- **Missing data mechanisms** - The explanation of the mechanisms is a very approachable and non-technical. However, the book provides an in depth, at least compared to other resources, guide on how to test for MCAR mechanism.

- **Traditional missing data handling strategies** - Chapter 2 is entirely dedicated to describing traditional methods for dealing with missing data methods and 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 (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. The imputation phase is described according to the **data augmentation algorithm**[1]. Great attention is payed to the Bayesian nature of this algorithm (there is even an introductory chapter on Bayesian statistics). The pooling phase sections describes the pooling of point estimates (section 8.3) and standard errors (section 8.5), along with multiple parameters significance testings (D1, D2, etc. [see sections 8.11 through 8.13]).

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 a good introduction to Multiple Imputation through Chained Equation and a technical companion to the R package that implements the MICE algorithm.

- Missing data mechanisms are described through probability models for a missing data indicator variable, as introduced by Rubin (1976) (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 discussed, and very little space is dedicated to traditional methods on ad hoc solutions [see section 1.3]), and in particular to the MICE algorithm.
  Multiple Imputation is discussed first in the case of univariate missingness (ch. 3) and then extended to a multivariate scenario (ch. 4). In particular:

[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 with how van Buuren (2012) used it. Pay attention to this in future readings

  - Chapter 3 focuses on the ability of MI to include not only noise around the prediction but also model (parameters) uncertainty. 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], and predictive mean matching and classification and regression trees are discussed as well.
  - Chapter 4 focuses on multiple imputation techniques that deal with multivariate missing data. The main difference with respect to univariate cases is that multiple patterns of missingness are possible when more variables are missing a value. First, the chapter describes the types of missing data patterns, and describes how to use the MICE package to detect/describe these patterns. Then, it discusses the three main approaches to multivariate missing data imputation: monotone data imputation, joint modelling (**JM**), and full conditional specification (**FCS**). FCS is the preferred approach by the book.
- **Analysis** and **pooling phase** are dealt with in chapter 5.
- In chapter 6, **diagnostic techniques** for multiple imputation are tackled. In particular, great attention is devoted to:
  - algorithmic convergence diagnostic (section 6.5.2)
  - and 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 (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.

This article provides a concise mathematical workout of Maximum Likelihood estimation of multivariate normal models with a monotone missing data pattern (from bivariate to any-variate generalisation).

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

The EM algorithm is introduced in this paper as an approach to iteratively compute maximum likelihood estimates when data is incomplete.

The algorithm is introduced first through a numerical example involving discrete variables, then, formally, for the case in which the complete data distribution belongs to the exponential family. Finally, it is extended for any distributional family. The article extensively discusses the general properties of the EM algorithm with great attention to the technical details. In section 3, the authors provide the mathematical justification for the effectiveness of the algorithm in finding the maximum likelihood estimate of a vector parameter when the complete-data likelihood is unkwon/intractable. Finally, the application of the EM algorithm is discursively exemplified in different scenarios such as missing data, and mixture modelling.

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 authors provide a technical description of the *Missing Information Principle*, as defined through the decomposition of the information 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 hypothetical 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.

The article defines the conditions under which statistical inference can still be considered proper while ignoring the missing data mechanisms. This is the reference source for the technical definitions of data Missing at Random (MAR), Observed at Random (OAR), and distinctness of the model parameter of interest (i.e. the object of inference) and the nuisance parameter (i.e. the parameter of the missing data process). Apart from rigorous theorems the article presents easy-to-grasp examples of what MAR, OAR, and distinctness mean and imply.

Rubin, D. B. (1978). Multiple imputations in sample surveys – A phenomenological Bayesian approach to nonresponse. , *1*, 20–34.

The article defines the foundations of Multiple Imputation, the procedure that imputes missing values reflecting the uncertainty within an imputation model and the sensitivity of inferences to different imputation models. The article distinguishes three fundamental tasks in the process of creating imputations: a modelling task, that chooses a model for the data, an estimation task, that finds the posterior distribution for its model parameters, and the imputation task that takes draws from the associated predictive distribution given the observed data.

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.

> Raghunathan *et al* introduce the SRMI (Sequential Regression Multivariate Imputation) approach to missing data imputation. The approach imputes the missing values on a variable-by-variable basis, by using posterior predictive distributions of the missing data, conditional on the observed data.
>
> Apart from describing the principles, strengths, and weakness of the approach, their work also provides detailed instructions on how to perform multiple 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 approach performs 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.

> Different algorithms to perform Multiple Imputation by FCS, according to the type of dependent variables, 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 conditional distributions is given. A good technical complement for an interested read is Arnold (2001). Simulations and study cases are then discussed in detail for both univariate and multivariate missing data scenarios, accounting for different variable types (i.e. continuous, 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.
>
> Finally, a simulation study is performed to assess the consequences of (in)compatibility. This is just an exemplifying scenario: not all possible incompatibility schemes can be considered. Yet, it does provide compelling evidence in favour 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.

> Andrige and Little point out two main advantages of hot deck imputation methods: it results in a rectangular data and it does not rely on any model specification. Different measures and methods for creating the donor pool are reviewed along with how to account for missing data patterns (monotone or *Swiss cheese*) and how to incorporate sampling weights.

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

> In this work, Rubin reviews the state of multiple imputation 18 years after the publication of his seminal works in 1976 and 1978. This paper clarifies the two main goals that Multiple Imputation was designed to achieve: 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 scientific estimand.
>
> The concept of statistical validity is operationalised with clarity by distinguishing between: randomisation validity and confidence validity.
>
> Furthermore, the concept of *proper* multiple imputation is discussed by summarising its main requirements.
>
> Finally some criticisms to MI are addressed through the lenses provided by 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 comparative guidelines and recommendations on different aspects of MI:
>
> - Handling different dependent variable distributions. In particular, there is a helpful section dealing with how to handle imputation for skewed continuous variables.
> - The question of which variables to include is addressed along with the issue of preserving all the relationships included in the analysis model, when defining the imputation model. Three main ways of dealing with this issues are presented (i.e. passive approach, improved passive approach using PMM, and JAV).
> - Number of imputations – The authors criticise the *efficiency argument* (that usually leads to the rule of thumb $m = 5$ is adequate for FMI $\leq .25$) through the *replicability argument*, centring the discussion around the decision of m in terms of Monte Carlo errors. The author'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

Belin, T. R., Hu, M.-Y., Young, A. S., & Grusky, O. (1999). Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine*, *18*(22), 3123–3135.

The General Location Model can be used for imputation for mixed data types (categorical and continuous) according to a data augmentation algorithm (see Schafer, 1997, ch. 9). This paper shows how this DA-GLM approach becomes exceedingly complex as the number of variables increases: for a dataset with 16 binary and 18 continuous variables the saturated model has more than 1 million parameters.

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 variables 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 variables based on their correlation with the variable(s) of interest that presents missing values, and the missingness mechanism). Through 4 different simulation setups, the authors study the consequences of including or excluding auxiliary variables, from each of these categories, and ultimately provide compelling evidence to prefer an inclusive strategy.

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) software implementations of ML methods do no facilitate it. This topic is thoroughly explored by Graham (2003).

A final point stressed by the article is that the consequences of the inclusive or restrictive strategies heavily depend on the type of MAR. For example, in their simulations, Collins *et al* found that the estimated mean of a variable Y of interest was biased when the auxiliary variable Z was excluded from the data handling model, if the probability of missingess was linearly related to Z (MAR-linear); however, that was not the case when the probability of missigness was larger for extreme values of Z or a function of the correlation between Z and Y.

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 shows, through multiple simulation studies, that including auxiliary variables, when using a FIML (Full Information Maximum Likelihood) approach to handling and analysing datasets with missing data, can be done relatively easily under the structural equation modelling framework.

With this paper, Graham introduces the *Saturated Correlated model* and *the extra dv model* to include auxiliary variables in a SEM model. Ultimately, he showed that it is possible to include auxiliary variables 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 of a substantive model that does not include them).

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 insightful 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 this principle, "information available from an incomplete data set is equal to complete information minus missing information". This implies that the fraction of missing information can be thought of as the ratio of missing information over the complete information. The definition of the Fraction of Missing Information (FMI) is explored under both the Maximum Likelihood and the Multiple Imputation framework. The fundamental contribution of this paper is in fact showing how FMI is not a prerogative of MI.

A final contribution is the 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.

Transformed variables such as squared and interaction terms need special attention when Multiple Imputation is performed. This article compares two main methods to impute them: the *transform then impute* and the *impute then transform* approach.

Through extreme missing data scenarios (100% missignes), and example data analyses, the author shows that the *transform then impute* method is the one that best preserves the mean and covariance structure of the original data and provides unbiased point estimates of regression estimates. As a result, he strongly recommends such method.

Some variants of these methods are included in the comparison, namely *passive imputation*, a more sophisticated but equally flawed version of the

impute then transform approach, and *stratify, then impute* for interactions between categorical and a continuous variables.

The main setup considered in this study is that of a linear regression model fitted to a dataset with MAR missing data. However, the author addresses the extension of the claims to models for binary dependent variables.