

Missing Data Analysis

Advanced Introductory Reading List

Compiled by Kyle M. Lang

2019-11-04

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

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*.
- 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).

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.
- 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 multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064.

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.

- 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.

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. *Psychological Methods*, 6(4), 330–351. doi: 10.1037//1082-989X.6.4.330
- 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
- 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.
- Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1), 265–291.