



Software for Multiple Imputation

Donald B. Rubin

Missing data are a pervasive problem in almost all areas of empirical research. It is now nearly a quarter century since multiple imputation was first proposed as a general solution to the problem of missing data. The first formally published seeds were in Rubin (JASA, 1977, p. 539): "One can think of the method given here as simply summarizing the results of simulations, where one uses the respondents to generate 'reasonable' hypothetical responses for nonrespondents." During this period of time, multiple imputation has become a major applied approach to the general problem of obtaining valid statistical inferences when faced with missing data, and may become the dominant approach in practice in the near future. A recent article with many references is Rubin (JASA, 1996).

One reason for this growth of multiple imputation is its suitability for modern computing environments, with the division of the two tasks of creating and analyzing a multiply-imputed data set. The task of analyzing an already-created multiply-imputed data set is relatively easy, requiring only (a) standard complete-data statistical routines to be applied to each complete data set created by the imputations, and (b) entirely general purpose macros for combining the repeated complete-data analyses to reach one inference. The task of creating a multiply-imputed data set is far more demanding - here is where modern computing is really essential. At present, there are two general approaches available, with accompanying software, for creating multiply-imputed data sets, both of which have advantages and limitations, which I briefly discuss here in anticipation of generating some discussion.

The first approach is a theme for SOLAS 2 and is particularly suited, though not limited to, cases with monotone patterns of missingness: when such a pattern occurs, fully principled and flexible statistical modeling for imputation can be done without extraneous assumptions (Little and Rubin, 1987, Chapter 6; Rubin, 1987, Section 5.4). A monotone pattern of missingness arises when the data matrix can be divided into observed and missing parts with a "staircase" line dividing them - e.g., variable 1 is observed on all the units (1-100), variables 2 and 3 are observed on units 1-75, but missing on units 76-100; variable 4 is observed on units 1-50 but missing on units 51-100. The great statistical advantage of a monotone pattern is that the creation of multiple imputations in a multivariate data set is reduced to a series of single variable imputations, which allows tremendous modeling flexibility. In principle, with a monotone pattern, we have all the flexibility that we have with complete data modeling (e.g., transformations of the outcome variable and predictor variables, interactions of predictor variables, nonlinear regressions [even for conditional variances], etc.). There is also a great computational advantage with a monotone pattern because the existence of missing data does not impose any need for iterative algorithms, either for maximum likelihood or Bayesian posterior estimation.

Of course, the problem in practice is that it is rare that any real data set has a perfectly monotone pattern of missingness. As a result, to retain all the great advantages of the monotone pattern, in practice, something less than fully principled often needs to be done to fill in missing values to create a monotone pattern. This suggestion to create a monotone pattern by filling in values (or even deleting observed values, which SOLAS 2 does not do) is, in fact, even older than the idea of multiple imputation (or EM -Dempster, Laird, Rubin, JRSS-B, 1977, for that matter); it appears in Rubin, 1974, JASA, Section 6.2. Moreover, the recognition of the importance of monotone patterns predates Rubin (1974); in fact, this article can be viewed as full-fledged extension of Anderson (1957, JASA), which itself was a neat derivation of maximum likelihood estimates for the simple case of bivariate normal monotone missing data done first by Wilks (1932, Annals of Mathematical Statistics). An early application of this idea of reducing a missing data problem to one with a monotone pattern is given in Marini, Olsen and Rubin (1980, Sociological Methodology).



The way SOLAS 2 gets to a monotone pattern is by first sorting to get close to a monotone pattern and then multiply-imputing the values that destroy the monotone pattern using a series of carefully selected available-case regressions. Other options are available in SOLAS 2, and these could be especially useful when much is known about the reasons for missing values. There is some "art" needed to do this reduction to a monotone pattern well, but once accomplished, the principled flexibility for creating multiple imputations is very rewarding, especially if the extra percent of missing information due to the non-monotone missingness is not large.

The second currently available general method for creating multiple imputations is to specify one encompassing multivariate model for the entire data set (at least conditional on completely observed variables), and then to use fully principled likelihood/Bayesian techniques for analysis under that model. This generates a posterior distribution for the parameters of the model and a posterior predictive distribution for the missing values (given the model specifications and the observed data values).

The primary example of this approach is Joseph Schafer's freeware (Schafer:www) based on Schafer (1997), which involves iterative Markov chain Monte Carlo computations. A more limited version is Gary King's freeware (King:www), which involves iterative maximization using EM and draws under an asymptotic normal approximation refined using SIR (Rubin, JASA, 1987). The theoretical advantage of this second general approach is that it is theoretically correct, no matter what the pattern of missing data, IF the specified model is correct AND the Bayesian/likelihood analysis method is correctly implemented.

The practical disadvantages of this second general approach, unfortunately, are just as clear as its theoretical advantages. First, despite the availability of some collections of multivariate models (e.g., the normal general location in Schafer's software), these collections are extremely limited - at least limited relative to the enormous collection of models available with the monotone-missingness approach. This limitation can be a serious concern in practice unless the person doing the imputation is a real expert, not only at statistical modeling with complete data and missing data, but also at making the software work in clever and often unanticipated ways. The second major disadvantage of using the "one encompassing model" approach is that, even if one of the limited models that is available in the software is appropriate for the data at hand, making an iterative program work can be a nightmare in practice, especially without the resources of Ph.D. level experts and the patience to deal with potentially misleading "nonconvergent" converging Markov chain Monte Carlo output or to assess the propriety of large sample asymptotic approximations refined with SIR, which despite my earlier hopes, may not work very well in large dimensions, especially when applied to large sample approximations where the underlying likelihoods can be multi-modal.

My own assessment is that unless a user has the resources to use Schafer's software with data that conforms pretty well to the underlying assumptions (or the statistical resources to enhance those specifications), the iterative versions of software for creating multiple imputations are not yet ready for reliable applications by the typical user. At this point, I think that the SOLAS 2 attack of pushing the data to be monotone and then being fully principled and flexible is a safer path for the typical user. This is a subjective assessment, which could be studied in real applications. Also, of course, there is the important consideration of having a validated and supported computing environment that is nearly always more associated with commercial products than freeware.

The future, of course, holds the promise of methods that combine the best features of both general approaches. For example, use the SOLAS 2 type of routine to sort the data matrix as closely as possible to conform to a monotone pattern; use Schafer's type of software (or perhaps the "incompatible Gibbs" software of T.E. Rathunathan at the Institute for Social Research (Ragunathan:www) to multiply-impute the missing values that destroy the monotone pattern; and



then use the SOLAS 2 type of routine to multiply-impute the monotonely missing values. Or perhaps even better, iterate the entire process, including the filling-in of the monotonely missing values. This latter approach describes one I am currently directing with a major data set of substantial importance, using a very knowledgeable third party statistical consulting company, but this effort is extremely time-consuming to set up and very expensive to implement (well into six figures and counting), and the creating of just one imputed data set takes more than week of CPU time on a modern minicomputer and days of diagnostic checking - clearly this approach is not yet ready for the average user even with today's fastest PCs.

Lest all these comments make the task of multiple imputation appear too daunting to attempt, I must add that even doing multiple imputation relatively crudely, using simple methods, is very likely to be inferentially far superior to any other equally easy method to implement (e.g., complete-cases, available cases, single imputation, LVCF) because the multiple copies of the data set allow the uncertainty about the values of the missing data to be incorporated into the final inferences; Heitjan and Rubin (JASA, 1990) provides some evidence for this statement, as does Raghunathan and Paulin (1998). And the use of SOLAS 2 (or Schafer's software or King's) to create a multiply-imputed data set is nearly certain to be far superior to any other generally feasible approach.



References

- Anderson, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52, 200-203.
- Dempster, A.P., Laird, N. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1, 1-38, with discussion and reply.
- Heitjan, D.F. and Rubin, D.B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85, 410, 304-314.
- Gary King, James Honaker, Anne Joseph, and Kenneth Scheve. 2000. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." Paper Presented at the annual meetings of the American Political Science Association, Boston. Current version available at <http://gking.harvard.edu/>
- Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons. Translated into Russian in 1991: Finansy and Statistika Publishers, Moscow, Andrei Nikiforev, translator.
- Marini, M.M., Olsen, A.R. and Rubin, D.B. (1980). Maximum likelihood estimation in panel studies with missing data. *Sociological Methodology* 1980, Chapter 11, 314-357.
- Raghunathan, T.E. and Paulin, G.D. (1998). Multiple imputation in the Consumer Expenditure Survey: evaluation of statistical inference. *Proceedings of the Business and Economics Section of the American Statistical Association*, 1-10.
- Raghunathan, T.E.: www.isr.umich.edu/src/smp/ive
- Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, 69, 346, 467-474, Section 6.2.
- Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 359, 538-543.
- Rubin, D.B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. Discussion of "The calculation of posterior distributions by data augmentation" by Tanner and Wong. *Journal of the American Statistical Association*, 82, 543-546.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 434, 473-489, with discussion 507-515, rejoinder 515-517, and extensive references 486-489.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
- Schafer, J.L.: www.stat.psu.edu/~jls
- Wilks, S.S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, 2, 163-195.