Journal of Statistical Software

December 2011, Volume 45, Issue 1. *http://www.jstatsoft.org/*

State of the Multiple Imputation Software

Recai M. Yucel
University at Albany, SUNY

Abstract

Owing to its practicality as well as strong inferential properties, multiple imputation has been increasingly popular in the analysis of incomplete data. Methods that are not only computationally elegant but also applicable in wide spectrum of statistical incomplete data problems have also been increasingly implemented in a numerous computing environments. Unfortunately, however, the speed of this development has not been replicated in reaching to "sophisticated" users. While the researchers have been quite successful in developing the underlying software, documentation in a style that would be most reachable to the greater scientific society has been lacking. The main goal of this special volume is to close this gap by articles that illustrate these software developments. Here I provide a brief history of multiple imputation and relevant software and highlight the contents of the contributions. Potential directions for the future of the software development is also provided.

Keywords: missing data, multiple imputation, software, computational algorithm.

1. Background

Methods specifically targeting missing values in a wide spectrum of statistical analyses are now part of serious statistical thinking due to many advances in computational statistics and increased awareness among sophisticated consumers of statistics. In particular, since its introduction by Rubin in 1976, inference by multiple imputation (MI) has been increasingly popular among both statisticians and practitioners. Following the seminal books by Rubin (1987) and Schafer (1997), MI has been and continues to be developed theoretically and adapted and implemented in in numerous statistical problems such as complex sample designs, high dimensional genetic data or measurement error (Reiter and Raghunathan 2007; Foulkes, Yucel, and Li 2008; Yucel and Zaslavsky 2005).

The key feature of MI over the other methods that are either parametric (e.g., likelihood-based) or non-parametric (e.g., weighting-based) is its versatility in the post-imputation phase as MI can serve multiple analytical goals using the same multiply-imputed datasets. Regardless of the nature of the post-imputation phase, MI inference treats missing data as an explicit source of random variability and the uncertainty induced by this is explicitly incorporated into the overall uncertainty measures of the underlying inferential process. This is accomplished by repeating the same complete-data analysis on the imputed data, and combining the estimates and standard errors under rules defined by Rubin (1987), including an explicit estimate of the degree of uncertainty due to the missing-data methodology.

To produce the imputations, some assumptions about the data (typically a parametric model) and the mechanism producing missing data need to be made. The assumed data model should be plausible and should be somewhat related to the analyst's investigation (Meng 1994). This model forms the basis to approximate the distribution in which the missing data conditional on observed data (i.e. predictive distribution of missing data). The software given in this volume develops computational routines for sampling from this approximate predictive distribution of missing data. The collection of the software presented here adapts roughly three approaches for the sampling.

The first approach jointly models variables subject to missingness, thus jointly samples from the underlying predictive distribution. The software by Yuan (2011) and Carpenter, Goldstein, and Kenward (2011) implement this approach. The second approach has been an excellent alternative to this method. It approximates the joint modeling approach with potentially incoherent variable-by-variable approach (Su, Gelman, Hill, and Yajima 2011; Buuren and Groothuis-Oudshoorn 2011; Royston and White 2011). While "incoherence" has been a subject of debate, this method has been quite successfully applied in many survey settings where joint approach is essentially not applicable. The final approach pertains to a collection of resampling-based and/or matching-based algorithms (Honaker, King, and Blackwell 2011; Yuan 2011).

Successful and thorough development of the MI paradigm allowed many of the current studies in social, medical or educational studies to consider MI. However, the lack of thorough software development as well as appropriate documentation (e.g., user-friendly manuals) has prevented a full incorporation of MI into these subject-matter fields. Most of the current software has only been developed since 2000s. Stand-alone Windows software **norm** (accompanying Schafer 1997), operating under a multivariate normality assumption, was arguably the first one available as both stand-alone software and S-PLUS library (Insightful Corp. 2003). Additional S-PLUS libraries for categorical and a mixture of categorical and continuous incomplete data later evolved into the engine of the bigger S-PLUS library called **missing** (Schimert, Schafer, Hesterberg, Fraley, and Clarkson 2000). In 2002, SAS (SAS Institute Inc. 2003) implemented some of Schafer's normal-based routines along with other routines for monotonic missingness patterns as well as matching-based MI routines (`PROC MI` and `PROC MIANALYZE`). This initiation by SAS was an important step in making the modern missing data techniques accessible and visible to a wider group of practitioners. Other major implementations of MI routines are found in **IVEware** (Raghunathan, Lepkowski, and VanHoewyk 2001) and **mice** (Buuren and Groothuis-Oudshoorn 2011) which both provide alternative imputation strategies for survey data with variables measured on a diverse set of scales. **mice** has been implemented as both S-PLUS library and an R package (R Development Core Team 2011) while **IVEware** is a SAS macro (SAS Institute Inc. 2003).

1.1. Goal of the special volume

Numerous substantive reasons have motivated statistical developers to expand and relax the assumptions that have been inherent to MI in simple settings (e.g., allowing multilevel structures in the process of MI). This special volume will provide a unique forum on the current state of MI software targeting a wide audience including sophisticated consumers of statistics and statistical researchers. The clarity of the individual manuscripts with illustrations on the underlying software using real datasets will not only contribute to the literature of statistical software development, but it will also disseminate the current state-of-the-art MI software to researchers in the subject-matter fields.

The overall goal is to provide a comprehensive volume on the MI software providing practitioners a much needed source that is solely focused on the most recent software developments and their illustrations. Further, diagnostic tools provided in this volume (Su *et al.* 2011) allow practitioners to assess the quality and validity of imputations, which have been missing from arguably all MI software up to date.

2. Summary of the special volume

Manuscripts are organized following the underlying "imputation" philosophy implemented by the respective software. First group shares the common theme of variable-by-variable approach (also referred as chained imputation models). This approach is particularly useful in problems with a set of incompletely-observed variables with diverse set of measurement scales (e.g., continuous, categorical, count and semi-continuous) and in problems complicated by common survey practices including skip patterns and truncation. First paper in this group is by Su *et al.* (2011). Their software implements flexible imputation techniques via chained imputation models and diagnostic tools that allow users to assess plausibility of the assumed imputation models. Specifically, their package **mi** features flexible choice of predictors, models, and transformations for chained imputation models; binned residual plots for checking the fit of the conditional distributions used for imputation; and plots for comparing the distributions of observed and imputed data in one and two dimensions. Bayesian models are also used to construct more stable estimates when data are sparse and supported by a prior knowledge.

The second contribution is by Buuren and Groothuis-Oudshoorn (2011) illustrating an increasingly popular approach to producing multiple imputations in settings pertaining to variables that are of varying natures and measured with restrictions. They present the most recent version of their R (R Development Core Team 2011) package called **mice** which imputes incomplete values by fully conditional specification. This package offers many practical solutions including predictor selection, passive imputation and automatic pooling to combine estimates from the multiply imputed datasets. These features are also extended to the multilevel continuous data. Finally, this version adds a capability of multilevel MI and interactive use with SPSS (IBM Corporation 2011). The third contribution presents an implementation of a similar approach in Stata (StataCorp. 2011). The manuscript by Royston and White (2011) describes **ice** which is the Stata module of the approach using the fully automatic pooling to produce multiple imputation. Royston and White (2011) illustrate this fully-integrated module in Stata using real data from an observational study in ovarian cancer.

Joint modeling approach follows the variable-by-variable approach. Carpenter and his colleagues describe a comprehensive module called **REALCOM-IMPUTE** of the multilevel model fitting software **MLwiN** (Carpenter *et al.* 2011). Variables subject to missing values are modeled under a multivariate latent normal model with random-effects, which is used as a basis to approximate the underlying posterior predictive distribution. The authors use Markov chain Monte Carlo (MCMC) simulation techniques to fit the imputation models and thus draw the multiple imputations. The software also allows for weights to account for sampling design both at level 1 and level 2. A variety of variables can be imputed: continuous, ordinal or nominal. Users can further analyze the imputed datasets under multilevel models and combine estimates using MI rules defined by Rubin (1987).

Another increasingly popular package is PROC MI and PROC MIANALYZE procedures of SAS. Yuan (2011) illustrates how to conduct MI inference in SAS. PROC MI implements three major techniques one can adopt to produce multiple imputations. Specific choice of these techniques depends on the missingness pattern and the type of imputed variable. For the problems with monotone patterns of missingness (i.e. a variable missing implies that all subsequent variables to be missing), one can choose from the following three methods depending on the type of the variable(s) to be imputed: matching (using propensity score or predictive mean) or MCMC which draws imputations from a multivariate normal if the underlying variables are continuous. If they are categorical, one can choose logistic regression or discriminant-function-based method to match. For the arbitrary patterns of missingness, one would have to approximate the underlying posterior predictive distribution using a multivariate normal distribution with a set of priors provided by PROC MI (e.g., ridge or Jeffreys prior).

The final contribution illustrates **Amelia** (Honaker *et al.* 2011). **Amelia** integrates two important computational tools EM and bootstrap to produce multiple imputations (Dempster, Laird, and Rubin 1977; Efron 1979). It implements a new computationally-improved EM-bootstrapping algorithm as an alternative to MCMC-based solutions. The imputation model still relies on a joint model, but the underlying sampling from the posterior predictive distribution is fundamentally different. Because the computations are centered around maximum likelihood (or posterior mode) estimates and it merely uses a re-sampling-based algorithm, it provides a computational efficiency. It also includes features to accurately impute cross-sectional datasets, individual time series, or sets of time series for different cross-sections. Finally, it allows users to facilitate graphical diagnostics for the imputed datasets.

3. Discussion and outloook

Software development is one of the significant keys to dissemination of the statistical methods. Without it, the greater scientific community simply does not have the means to access to state-of-the-art techniques. Due to high prevalence of missing data in research problems relying on empirical evidence, it is critical for the statistical community to provide objective and open source for missing data software. This special volume aims to provide exactly this, and it is my hope to see updates to this special volume to provide statistical and substantive literatures with the up-to-date documentation of software. The diversity of the contributions to this special volume provides an impression about the progress of the last decade in the software development in the multiple imputation.

It should be noted that this volume is not intended to be the exclusive source of the multiple imputation software. There are many other software companion to the methods developed so far. Some of the most commonly-used software include R packages **Hmsic** (Harrell 2011, function `aregImpute`), **norm** (Novo and Schafer 2010), **cat** (Harding, Tusell, and Schafer 2011), **mix** (Schafer 2010) for a variety of techniques to create multiple imputations in continuous, categorical or mixture of continuous and categorical datasets. Another useful R package for imputing continuous variables in clustered or longitudinal designs is **pan** (Schafer 2011; Schafer and Yucel 2002). There is also a very important package in the form of SAS macro for multiple imputation using a sequences of regression models. This SAS-callable program is called **IVEware** written by Raghunathan *et al.* (2001) and very similar to the R and Stata implementation of **mice** and **ice**.

The implemented methodology of MI has so far focused on the improved computational algorithms geared towards relatively simpler data designs. In other words, software development for MI is just starting. There are many problems for which the greater scientific community is looking for principled and ready-to-use tools. Some examples include extensions of variable-by-variable-based methodology to clustered designs, multilevel datasets, incorporation of non-ignorable mechanisms. I believe that the software development will be greatly helped by open-source forums such as R as it provides a great forum for steady improvements via users' feedback and constructive criticism.

4. Acknowledgments

As the editor of this special volume, I would like to express my gratitude to all authors who submitted contributions. I am in particular very thankful for the valuable and extremely constructive work of the reviewers.

References

Buuren SV, Groothuis-Oudshoorn K (2011). "**mice**: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(3), 1–67. URL http://www.jstatsoft.org/v45/i03/.

Carpenter JR, Goldstein H, Kenward MG (2011). "**REALCOM-IMPUTE** Software for Multilevel Multiple Imputation with Mixed Response Types." *Journal of Statistical Software*, **45**(5), 1–14. URL http://www.jstatsoft.org/v45/i05/.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society B*, **39**, 1–38.

Efron B (1979). "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics*, **7**, 1–26.

Foulkes AS, Yucel RM, Li X (2008). "Mixed Modelling with Ambiguous Clusters: A Likelihood-Based Approach." *Biostatistics*, **9**, 635–657.

Harding T, Tusell F, Schafer JL (2011). ***cat**: Analysis of Categorical-Variable Datasets with Missing Values*. R package version 0.0-6.3, URL http://CRAN.R-project.org/package=cat.

Harrell FE (2011). ***Hmisc**: Harrell Miscellaneous.* R package version 3.9-0, URL http://CRAN.R-project.org/package=Hmisc.

Honaker J, King G, Blackwell M (2011). "**Amelia** II: A Program for Missing Data." *Journal of Statistical Software*, **45**(7), 1–47. URL http://www.jstatsoft.org/v45/i07/.

IBM Corporation (2011). *IBM SPSS Statistics 20.* IBM Corporation, Armonk, NY. URL http://www-01.ibm.com/software/analytics/spss/.

Insightful Corp (2003). *S-PLUS Version 6.2.* Seattle, WA. URL http://www.insightful.com/.

Meng XL (1994). "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science*, **10**, 538–573.

Novo AA, Schafer JL (2010). ***norm**: Analysis of Multivariate Normal Datasets with Missing Values.* R package version 1.0-9.2, URL http://CRAN.R-project.org/package=norm.

SAS Institute Inc (2003). *SAS/STAT Software, Version 9.1.* Cary, NC. URL http://www.sas.com/.

Raghunathan TE, Lepkowski JM, VanHoewyk J (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology*, **27**, 1–20.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reiter JP, Raghunathan TE (2007). "The Multiple Adaptations of Multiple Imputations." *Journal of the American Statistical Association*, **102**, 1462–1471.

Royston P, White IR (2011). "Multiple Imputation by Chained Equations (MICE): Implementation in Stata." *Journal of Statistical Software*, **45**(4), 1–20. URL http://www.jstatsoft.org/v45/i04/.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley and Sons, New York.

Schafer JL (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Schafer JL (2010). ***mix**: Estimation/multiple Imputation for Mixed Categorical and Continuous Data.* R package version 1.0-8, URL http://CRAN.R-project.org/package=mix.

Schafer JL (2011). ***pan**: Multiple Imputation for Multivariate Panel or Clustered Data.* R package version 0.3, URL http://CRAN.R-project.org/package=pan.

Schafer JL, Yucel RM (2002). "Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values." *Journal of Computational and Graphical Statistics*, **11**(2), 421–442.

Schimert J, Schafer JL, Hesterberg T, Fraley C, Clarkson DB (2000). *Analyzing Data with Missing Values in S-PLUS*. Data Analysis Products Division, Insightful Corporation, Seattle, WA.

StataCorp (2011). *Stata Data Analysis Statistical Software: Release 12*. StataCorp LP, College Station, TX. URL http://www.stata.com/.

Su YS, Gelman A, Hill J, Yajima M (2011). "Multiple Imputation with Diagnostics (**mi**) in R: Opening Windows into the Black Box." *Journal of Statistical Software*, **45**(2), 1–31. URL http://www.jstatsoft.org/v45/i02/.

Yuan Y (2011). "Multiple Imputation Using SAS Software." *Journal of Statistical Software*, **45**(6), 1–25. URL http://www.jstatsoft.org/v45/i06/.

Yucel RM, Zaslavsky AM (2005). "Imputation of Binary Treatment Variables with Measurement Error in Administrative Data." *Journal of the American Statistical Association*, **100**, 1123–1132.

Affiliation:

Recai M. Yucel
Department of Epidemiology and Biostatistics
One University Place, Room 139
School of Public Health
University at Albany, SUNY
Rensselaer, NY 12144–3456, United States of America
E-mail: ryucel@albany.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 45, Issue 1 *Submitted:* 2011-09-19
December 2011 *Accepted:* 2011-11-20