Multiple Imputation via a Semi-Parametric

Probability Integral Transformation

BY

IRENE HELENOWSKI
B.A., Northwestern University, 1998
M.S., University of Wisconsin – Madison, 2001

THESIS
Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences
in the Graduate College of the
University of Illinois at Chicago, 2012

Chicago, Illinois

Defense Committee:

Hakan Demirtas, Chair and Advisor
Sally Freels
Donald Hedeker
Hua Yun Chen
Borko Jovanovic, Northwestern University

This thesis is dedicated to the memory of my father, Adam T. Helenowski, M.D. and to my mother, Irena Helenowski, D.D.S.

ACKNOWLEDGMENTS

This work was made possible through the supervision and collaboration of my dissertation committee. Firstly, I thank my advisor, Hakan Demirtas, PhD., for his advice, support, literature recommendations, computational recommendations, advice of presentation formatting, and countless hours of review. I thank Sally A. Freels, PhD., for her help in improving my writing and presentation techniques. I further appreciate the advice Donald R. Hedeker, PhD., gave regarding inclusion of real data examples and computational approaches. I am grateful to Hua Yun Chen, PhD., for his recommendations on several computational approaches, especially regarding the calculation of eCDF values. Finally, I thank my committee member and faculty colleague at Northwestern, Borko D. Jovanovic, PhD., also for constant review of my work, recommendations for presentation, and suggestions for real data examples.

I further thank my family, first and foremost my parents, Adam T. Helenowski, M.D., and Irena Helenowski, D.D.S., for always supporting me in my educational pursuits throughout my academic and professional careers. I also thank my sister, Joanna Helenowska-Gittler for helping me with my grade school science projects and introducing me to Dr. Harry Herzer from NASA, my brother, Tomasz K. Helenowski, M.D., for the gift of the laptop on which I ran simulations for countless hours and for his words of encouragement coming from his own experiences involving his work in stereotactic radiosurgery, and my other brother Jacek Helenowski, who has also influenced my interest in mathematics and the sciences through his work in cold cathode technology and the construction of his state-of-the-art LEED (Leadership in Energy and Environmental Design) home. Likewise, I acknowledge the support of my sister-in-law, Susan Helenowski, R.N., brother-in-law, John Gittler, and sister-in-law, Marta Helenowski, M.D. (who helped me understand the pathological terms and mechanisms involved in several of my real data examples), and of my nieces and nephews, Anna, Julia,

Mark, Christopher, Eryk, Alyssa, and Luxa.

I would also like to thank my boss, Alfred W. Rademaker, PhD., for allowing me to continue with this academic endeavor while working in my position as a statistical analyst in the Northwestern University Department of Preventive Medicine. I am very grateful and honored to have worked under his supervision for almost ten years at Northwestern. As mentioned before, I am also very grateful to my faculty colleague and member of my dissertation committee, Borko D. Jovanovic, PhD., as well as to fellow members of the Northwestern University Prostate SPORE team, Timothy M. Kuzel, M.D., and Robin G. Leikin, PhD, who have helped me attain permission for use of the Prostate SPORE database and to Aishwarya Parameswaran for providing me with the variables and general information from the database. I also appreciate the continued support of colleagues Joan Chmiel, PhD., Jungwha 'Julia' Lee, PhD, Kwang-Youn Kim, PhD, Joseph Kang, PhD, Mary Kwasny, PhD, Laura Colangelo, M.S., Veronica Jackson, and many others in the Biostatistics Division of Preventive Medicine at Northwestern.

Lastly, I thank my friends who provided me with love, encouragement, and humor during trying times of this endeavor including Jennifer Beaumont, M.S., Yoonsang Kim, Ph.D., Oksana Pugach, M.S., Beth Andrews, PhD., Chris Henry, PhD., Jason Fine, PhD., John Cursio, M.S., Chris Cho, M.S., Paul Duty PhD., Kush Kapur, PhD., Lisa Sromek, Erika Heier, Lynn Vance, Peggy Hu, and Jackie Bernardin, among countless others.

IBH

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF ABBREVIATIONS

MCAR	Missing Completely At Random
MAR	Missing At Random
eCDF	empirical Cumulative Distribution Function
PDF	Probability Distribution Function
PIT	Probability Integral distribution function Transformation
LGMI	Lurie-Goldberg Multiple Imputation algorithm
AE	Average Estimate of parameter obtained from imputed data
SB	Standardized Bias
RMSE	Root Mean-Square Error
CR	Coverage Rate
AW	Average Width of 95% confidence intervals for estimates from imputed data
CC	Complete-case analysis
SPORE	Specialized Program of Research Excellence
NYC HANES	New York City Health and Nutrition Examination Survey

SUMMARY

In this work, we propose approaches for imputing continuous, binary, and mixed data by first mapping these data to normally distributed values and then applying multiple imputation so that distributional assumptions for the original data can be relaxed. For continuous data, our technique incorporates transformations and back-transformations suggested by the Lurie and Goldberg (1998), and also involves calculating the marginal empirical cumulative distribution function (eCDF), instead of the cumulative distribution function of a specific distribution for each variable in the data. Using eCDF values is the key step in allowing us to carry out the imputation procedure while allowing us to relax parametric assumptions for the original data. For binary data, we employed methods presented in Emrich and Piedmonte (1991) and Demirtas and Doganay (2012) for generating multivariate binary data from an underlying normal distribution associated with tetrachoric correlations derived from the pairwise phi coefficients relating the variables of the binary data. Dichotomizing these normally distributed data using quantiles has the same role as computing eCDF values in the case for continuous data in that it allows for back-transforming imputed values while relaxing specific parametric assumptions. Additionally, our approach for imputing mixed data incorporating both the Lurie-Goldberg algorithm and eCDF computation with continuous data and the use of an underlying multivariate normal distribution and quantiles used in dichotomizing with binary data. Applying our method to simulated continuous data following the normal, *t*, or Gamma distributions and to simulated binary and mixed data led to promising results in both bivariate and multivariate cases. The approach also performed well with real data sets obtained from the NYC HANES and Prostate SPORE (Grant #: P50 Ca 090386) databases, as well as for simulated data resembling these real data sets. We conducted our simulation studies under assumptions of the MCAR mechanism. We believe that this approach will be a useful tool for investigators analyzing data with significant missing information.

Multiple Imputation via a Semi-parametric Probability Integral Transformation

Irene B. Helenowski, Ph.D.
Division of Epidemiology and Biostatistics
School of Public Health
University of Illinois at Chicago
Chicago, Illinois (2011)

Dissertation Chairperson: Hakan Demirtas, Ph.D.

In real data scenarios, the distribution of the data is often unknown. Therefore, methods for imputing data which relax distributional or model assumptions may be of great interest to investigators. Here, we propose semi-parametric approaches allowing us to relax distributional assumptions when imputing continuous data, multinomial or loglinear model assumptions when imputing binary data, and general location model assumptions when imputing mixed continuous and binary data. The nonparametric portion of our methods involves mapping data to normally distributed values via empirical cumulative distribution (eCDF) or quantile computation and the parametric portion involves multiple imputation under the normality assumption via joint modeling. Applying our approaches to data generated under the MCAR mechanism and to real data from databases of the Northwestern University SPORE in Prostate Cancer (Grant #: P50 Ca 090386) and New York City Health and Nutrition Survey gave promising results.

1. INTRODUCTION

One method currently employed by many investigators handling a substantial amount of missing data is multiple imputation. Multiple imputation has been extensively studied under assumptions of data following the normal distribution (Little,1992 ; Little and Rubin, 2002; Rubin, 1987; Rubin, 1996; Schafer,1997; Schafer, 1999; Yuan and Bentler, 2000), as well as for data following other distributions (Demirtas, 2007; Demirtas and Hedeker, 2007; Demirtas, 2008; Demirtas *et al*., 2008; Demirtas and Hedeker, 2008; Gold and Bentler, 2000; Gold *et al*., 2003). Nevertheless, the exact distribution of real data is not commonly known; thus, developing an imputation technique not dependent on the actual distribution of the original data is critical. First, we propose a method employing one algorithm presented in Lurie and Goldberg (1998) and an application of the eCDF (empirical cumulative distribution function) to impute continuous data without positing any specific distributional assumptions directly on that data. Our method involves mapping the data to normally distributed values, similar to those transformations in the Lurie-Goldberg algorithm, imputing missing values (Schafer, 1997), and then back-transforming the imputed data to the range of the originally observed data (Barton and Schruben, 1993). Inclusion of principles from the Lurie-Goldberg algorithm in our method leads to the conservation of relationships among the variables in the data. Furthermore, incorporating computation the eCDF values allows transformation and back-transformation without assuming specific marginal distributions.

Next, we propose a method using the principles from generating binary data using a normal distribution, as described in Emrich and Piedmonte (1991) and Demirtas and Doganay (2012), to impute binary data. We first impute the normally distributed values generated based on a mean of 0 and a correlation matrix based on pairwise tetrachoric correlation coefficients, where each coefficient is computed from the distribution and phi coefficient associated with the observed data in the corresponding pair of binary variables. We then dichotomize the normally distributed values based on quantiles corresponding to probabilities obtained from the distribution of the observed binary data. Finally, we incorporate both proposed approaches given here for imputing continuous and binary data to impute mixed continuous and binary data. Our approaches to impute continuous, binary, or mixed data can therefore aid analyses of data with significant missing information and relaxing assumptions about the specific parametric distribution that the observed data follow.

2. MULTIPLE IMPUTATION

2.1 Definition

Multiple imputation is an attractive approach for handling missing data. It is preferred to ad hoc methods such as complete-case and available-case analyses when the amount of missing values is substantial, as the latter approaches may lead to inefficiency and bias in the data analysis (Schafer, 1997). We define multiple imputation as a Markov chain Monte Carlo (MCMC) technique replacing missing values with plausible values from a predictive distribution, such as a Bayesian predictive distribution of parameters given the observed data.

2.2 Missing Data Mechanisms and Patterns

There are three types of missing data mechanisms, namely:

a. MCAR (Missing Completely at Random)

b. MAR (Missing at Random)

c. MNAR (Missing Not at Random)

To further explain these three mechanisms, we first introduce some conventionally used nomenclature, as discussed in Demirtas (2004). We denote Y_{com} as the complete data set, Y_{obs} as the values in the data set observed, and Y_{mis} as the missing values in the data set. We also define a vector R as an indicator for missingness, where

$$R = \begin{cases} 1, & Y \text{ is observed} \\ 0, & Y \text{ is missing} \end{cases} \qquad (2.2.1)$$

for $j = 1, \ldots, n$ observations.

Under MCAR, a special case of MAR, we assume that the probability of missingness does not depend on the observed or missing data, i.e.:

$$P(R = r \mid Y_{obs} = y_{obs}, y_{mis}; \delta) = P(R = r, \delta) \qquad (2.2.2)$$

where r and y_{obs} are realizations of R and Y_{obs}, respectively, and δ is the set of parameters for the conditional distribution of R given Y_{com}. Under MAR, on the other hand, the missingness depends on only the observed data, such that:

$$P(R = r \mid Y_{obs} = y_{obs}, y_{mis}; \delta) = P(R = r \mid Y_{obs} = y_{obs}, \delta) \qquad (2.2.3)$$

Finally, for MNAR, the missingness may depend on both the missing and observed values. We note that the MNAR mechanism is associated with non-ignorable missing data, since the probability of missingness depends on the missing values themselves (Rubin, 1987; Heitjan and Rubin, 1991; Glynn *et al*. 1993; Schafer and Olsen, 1998; Little and Rubin, 2002; Demirtas and Schafer, 2003; Demirtas, 2004; Demirtas, 2004; Demirtas, 2005).

We can also describe missing data in terms of patterns as univariate, monotone, and arbitrary. To define these patterns, we assume that we have a multivariate data set with k variables. In the univariate pattern, $k - 1$ variables are completely observed, while one variable has missing entries. In the monotone pattern, variables are ordered with respect to increasing fractions of missing information such that the $(j+1)^{th}$ to the k^{th} variable have the same missing fraction as the j^{th} variable for $j = 1, \ldots, k$, plus an additional amount of missingness. Lastly, in the arbitrary pattern, missing values can occur in any of the k variables at any entry (Schafer and Graham, 2002).

2.3 Multiple Imputation, Likelihood Based Methods, and Single Imputation

Several comparisons between multiple imputation and likelihood-based approaches, such as maximum likelihood, indicate the advantages of multiple imputation in situations with extensive missing data. Multiple imputation and maximum likelihood both rely on large sample approximations, but multiple imputation also incorporates the missing data mechanism. If values are missing at random, then multiple imputation and likelihood-based approaches tend to produce similar results under conditions outlined in Schafer (2003). With non-ignorable missingness, however, parameter estimation might not only depend on the observed data. Therefore, multiple imputation is a beneficial alternative for handling non-ignorable missing data (Collins *et al*., 2001; Schafer and Graham, 2002).

Multiple imputation provides more flexibility than maximum likelihood estimation due to the separation between the imputation and analysis components (Demirtas and Hedeker, 2008; Schafer and Graham, 2002). Imputation models will prove satisfactory for analysis if the imputer considers potential models that the analyst might fit to the data. Multiple imputation and maximum likelihood will lead to similar outcomes when the imputation and analysis models include the same parameters and are based on the same distributional assumptions. Results will also be similar when the imputation model is more general, i.e., contains more parameters than the analysis model, albeit standard errors for parameter estimates from the multiple imputation approach will be somewhat larger.

Multiple imputation methods, as well as likelihood-based methods, further differ from ad hoc methods, as case deletion and single imputation, since they treat the missing data as random values to be averaged over, whereas ad hoc methods modify the incomplete data to mirror a complete data set (Collins *et al*., 2001; Schafer and Graham, 2002). Multiple imputation is always preferred to single imputation, although single imputation may be acceptable when the fraction of data missing is less than

5% of the total data (Schafer, 1999). Schafer and Olsen (1998) note the advantages of multiple imputation over ad hoc methods in terms of including data uncertainty in summary statistics. Ad hoc methods do not account for such uncertainty and therefore can lead to distorted data distributions and relationships. Approaches such as single imputation underestimate the true variability of the parameter estimate by ignoring this uncertainty, leading to overestimated precision, inflated Type I errors, artificially narrow confidence intervals, and overly optimistic p-values (Schafer and Olsen, 1998).

Parameter estimation associated with multiple imputation involves averaging the Bayesian posterior distribution of our parameters over the conditional distribution of missing data given the observed data and can be described by the integral:

$$\int P(\theta \mid Y) P(Y_{mis} \mid Y_{obs}) dY_{mis} \tag{2.3.1}$$

We can derive the average estimate for the population quantity of interest, Q, as:

$$\overline{Q} = m^{-1} \sum \widehat{Q^{(j)}} \tag{2.3.2}$$

where $\widehat{Q^{(j)}}$ is the quantity obtained from the single imputation *j*. The total variance, T, of Q is given by:

$$T = \overline{U} + \left(1 + m^{-1}\right) B \tag{2.3.3}$$

where

$$\overline{U} = m^{-1} \sum U^{(j)} \tag{2.3.4}$$

is the average of the variance estimate for $Q^{(j)}$ and

$$B = (m-1)^{-1} \sum (Q^{(j)} - \overline{Q})^2 \tag{2.3.5}$$

Thus, *U* is the within-imputation variance and *B* is the between-imputation variance (Demirtas, 2004; Schafer and Olsen, 1998). *T* is incorporated into the approximation

$$T^{-1/2} \left(Q - \overline{Q}\right) \sim t_{\nu} \tag{2.3.6}$$

where the degrees of freedom ν is:

$$\nu = (m-1) \left[1 + \frac{\overline{U}}{\left(1 + m^{-1}\right) B} \right]^2$$

(2.3.7)

When $B >> U$, the fraction of missing data, λ, is large, the relative increase in variance due to nonresponse,

$$r = \frac{\left(1 + m^{-1}\right)B}{\bar{U}} \qquad (2.3.8)$$

is large, and the degrees of freedom is small, leading to inferential biases based on normal approximation. To improve the validity of normal approximation, increasing the number of imputations is therefore recommended (Schafer, 1997; Schafer and Olsen, 1998).

Rubin (1987) further derives the relative efficiency of a finite number of imputations, *m*, and an infinite number of imputations as:

$$(1 + \lambda / m)^{-1} \qquad (2.3.9)$$

where λ is the fraction of missing information. This derivation is based on the variance of the estimate in question conditional on the observed data. With this equation, we can see that, for example, with $m = 5$ imputations and 50% missing information, our relative efficiency is $(1 + 0.50/5)^{-1} = \frac{1}{1.1} = 91\%$ and with $m = 10$ imputations and 50% missing information, our relative efficiency is $(1 + 0.50/10)^{-1} = \frac{1}{1.05} = 95\%$. Thus, only five or ten imputations are sufficient for most analyses (Schafer, 1999). Schafer (1997) notes the number of imputations as a reason why the efficiency of multiple imputation is less than that of likelihood based approaches, since likelihood based approaches do not require $m > 1$ simulations. He also recommends estimates from maximum likelihood inference, for example, as a reference to be compared against estimates from simulation-based methods. The relative efficiency over an infinite number of imputations in (2.3.9) then reflects the relative efficiency of multiple imputation over maximum likelihood (Collins *et al*., 2001). Lastly, Schafer (1997) notes that determining the number of simulations in a Markov Chain Monte Carlo process, such as multiple imputation, requires accounting for an initial burn-in period and minimizing and stabilizing the Monte Carlo error.

2.4 Examples of Different Multiple Imputation Approaches

There are several examples of multiple imputation approaches for drawing values to fill in for missing data. In Section 2.6, we will focus on the approaches involving EM (Expectation-maximization) and data augmentation algorithms. Other approaches include hot deck imputation, Bayesian bootstrap (BB), and approximate Bayesian bootstrap (Rubin and Schenker, 1991). For example, hot deck imputation involves drawing values for imputation from the observed data with replacement with equal probability. Rubin and Schenker (1991) state that this approach leads to underestimated variance, however. The Bayesian or approximate Bayesian bootstrap approaches, on the other hand, are re-sampling algorithms where values are drawn from a population using probabilities based on an improper Dirichlet prior or a multinomial posterior distribution, respectively. Modifications to the approximate Bayesian bootstrap method have been introduced, although simulation studies have shown that these modified approaches are not necessarily superior to the original approximate Bayesian bootstrap (Demirtas *et al*., 2007).

Barnes *et al*. (2006) discuss regression-based multiple imputation methods, such as Bayesian Least Squares (BLS), predictive mean matching (PMM), and local random residuals (LRR). With Bayesian least squares, Bayesian regression is first used to derive a joint posterior distribution for the regression parameters. Parameters drawn from this distribution are used to derive the mean and covariance of the normal distribution from which the imputed values are then drawn. Predictive mean matching is also based on predictive values from regression analyses, but now the missing responses are filled in with actually observed values whose corresponding predicted responses are closest in value to the predicted responses of the missing entries. In local random residuals, values are selected from a pool of observed values closest to the predicted value for the missing entry. The residual for the selected observed value (i.e., the difference between the predicted and true value for the observed data) is then obtained and added to the predicted value for the missing entry.

Two other important multiple imputation approaches include joint modeling and chained equations. Joint modeling is based on the distribution:

$$\mathrm{P}(Y_{mis}|Y_{obs}, X, R) \tag{2.4.1}$$

for the data matrix $Y = \{Y_{mis}, Y_{obs}\}$ regressed on the matrix of covariates X and the R matrix containing the indicators of missingness in Y (Schafer, 1997). This approach has been implemented for multivariate normal data, discrete data in loglinear models, and mixed data, containing both categorical and continuous data (Schafer, 1999). Section 2.7 discusses the software packages written by Schafer (1997) used to implement these methods. Schafer (1997) notes that joint modeling provides a channel via multiple imputation for handling missing non-normal data, as well as normal data.

With joint modeling, Schafer and Olsen (1998) and Schafer (1997) discuss the use of regression based on the multivariate normal distribution for continuous data and regression involving the loglinear model, covered here in Section 2.6, for categorical data. For mixed data including both continuous and categorical variables, these works also suggest employing a general location model which incorporates both loglinear model and multivariate normal regression model components, likewise covered in Section 2.6. Lastly, a two-level linear regression model is implemented into the joint modeling approach when handling missing longitudinal data. These types of models are imposed on the complete data in order to impute values for missing entries. Schafer and Olsen (1998) mention that choosing a model for this approach is nontrivial, even when the model itself is robust, as in the case of applying the multivariate normal model to transformed data. For example, although a model under the normality assumption can be applied to ordinal or binary data, with the resulting imputed continuous values then rounded off to the nearest category, a loglinear model can be recommended as a preferable alternative.

Chained equations is a multiple imputation approach where each variable is imputed with a separate model conditional on all other variables. Van Buuren *et al*. (1999) summarize the goal of their multiple imputation approach with respect to handling missing blood pressure values in their study relating mortality to blood pressure, age, sex, and several other health factors. The authors first discuss applying linear regression imputation and including variables that would be used in complete-case analyses, variables that may have different distributions between observed and missing data, and variables explaining a substantial amount of variation in the variable to be imputed. They subsequently recommend omitting variables with too many missing entries from the final imputation model. The authors next review estimation of linear regression parameters in the imputation model and generation of new parameter estimates via drawing values from the derived posterior distribution of parameters. They then introduce chained equations by presenting Gibbs sampling with the conditional distributions for each variable to be imputed individually.

Missing entries are first filled in using random draws from the marginal distributions of the observed data. The first variable, Y_1, say, is next imputed conditional on the observed data and all other imputed data. The second variable, Y_2, say, is consequently imputed using all other data including the most recently imputed Y_1 values and so on until all incomplete variables are imputed. Van Buuren *et al.* (1999) note that this chained equation approach can also be extended to non-ignorable missing data by adjusting the parameters of the distribution from which the imputed values are drawn by a constant δ.

Van Buuren (2007) indicates how joint modeling and chained equations are related in some cases. For example, in the case of multivariate normal data, conditional densities constitute linear regression models with constant error variance and vice versa. Similarly, for binary data, there can be a logistic model for each variable as a response with the other variables as predictors. Furthermore, comparisons between joint modeling and chained equations indicate some advantages of chained equations over joint modeling due to greater flexibility in creating complicated multivariate normal models. Van Buuren *et al.* (1999) state another advantage of chain equations as requiring less iterations than other Monte Carlo Markov Chain techniques.

The chained equations approach nevertheless has some drawbacks if two conditional distributions, $P(Y_1 | Y_2)$ and $P(Y_2 | Y_1)$, for example, are incompatible, causing switching between isolated distributions, an outcome leading to ongoing research problems. For example, the number of iterations sufficient to stabilize the posterior distribution is still to be determined, as regression switching absorbs the uncertainty in the predictors of the model.

Many more approaches exist, including bootstrap approaches for drawing values from a frequentist perspective (Efron, 1994). As stated previously, we describe the incorporation of EM and DA into imputation methods under the multivariate normal distributional assumptions, discussed in Section 2.5.

2.5 Multiple Imputation under the Assumption of Normally Distributed Data

Multiple imputation methods are well-established under assumptions that the data following a normal distribution. Schafer (1997) presents situations where multiple imputation methods conducted under the normality assumption can also applied to non-normal data. Such situations involve transformations of the variables or linear functions conditional on normal residuals, where these latter functions can even be applied to discrete data. Schafer (1997) reviews the maximum likelihood estimation of parameters μ and Σ for normally distributed data given by

$$\bar{y} = n^{-1}\sum_{i=1}^{n} y_i \tag{2.5.1}$$

and

$$S = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T \tag{2.5.2}$$

for $i = 1, \ldots, n$ observations.

He then reviews the EM, or Expectation-Maximization, (Dempster *et al.*, 1977) and data augmentation (DA) (Tanner and Wong, 1987) algorithms that are implemented into the multiple imputation. The EM algorithm also has the attractive properties of providing good starting values and insight into convergence behavior (Demirtas, 2007; Demirtas *et al*., 2008). This algorithm can be described in terms of computing the conditional expectation of the complete and sufficient statistic, *T*, for the data (E-step) and then maximizing this expectation (M-step). For the multivariate normal distribution with $\theta = (\mu, \Sigma)$, *T* can be obtained via the maximum-likelihood estimation equations given in (2.5.1) and (2.5.2).

Next, data augmentation is performed. The augmentation involves two steps: the I-step (imputation) and the P-step (posterior). In the I-step, associated with drawing values $Y_{mis}^{(t+1)}$,

$$Y_{mis}^{(t+1)} \sim P(Y_{mis} \,|\, Y_{obs}, \theta^{(t)}) \qquad (2.5.3)$$

we can generate each entry *i* for Y^{t+1} from the distribution given Y_{obs} and the most current $\theta^{(t)}$

$$y_{i(mis)}^{(t+1)} \sim P(y_{i(mis)} \,|\, y_{i(obs)}, \theta^{(t)}) \qquad (2.5.4)$$

The P-step involves the posterior distribution:

$$\theta^{(t+1)} \sim P(\theta \,|\, Y_{obs}, Y_{mis}^{(t+1)}) \qquad (2.5.5)$$

where the posterior distribution for $\theta^{(t+1)}$ is updated using Y_{obs} and $Y_{mis}^{(t+1)}$.

For a univariate sample, i.e., for

$$y_i \sim \mathrm{N}(\mu, \sigma^2) \qquad (2.5.6)$$

with $i = 1, . . . , n$ observations, the posterior distributions would be given by

$$\mu \,|\, \sigma^2, Y_{obs} \sim \mathrm{N}(\bar{y}, n^{-1}\sigma^2) \qquad (2.5.7)$$

and

$$\sigma^2 \,|\, Y_{obs} \sim (n-1)S^2 \,/\, \chi^2_{n-1} \qquad (2.5.8)$$

where χ^2_{n-1} is a chi-square variate with n – 1 degrees of freedom (Schafer, 1999).

Under the assumption that our data follow the multivariate normal distribution, we can use Bayesian inference to obtain μ via the conditional distribution:

$$\mu \,|\, \Sigma \sim N(\mu_0, \tau^{-1}\Sigma) \qquad (2.5.9)$$

and Σ via

$$\Sigma \sim W^{-1}(m, \Lambda) \qquad (2.5.10)$$

where μ_0, τ, and Λ are fixed and known hyperparameters and W^{-1} is the inverse Wishart distribution. The two components of the Wishart and the inverted Wishart distributions are m and Λ, noted as the degrees of freedom and scale, respectively. The Wishart probability distribution function is proportional to:

$$P(Y \mid m, \Lambda) \propto |Y|^{\frac{m-k-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\Lambda^{-1}Y\right\} \qquad (2.5.11)$$

and the inverted Wishart probability distribution function is proportional to:

$$P(Y \mid m, \Lambda) \propto |Y|^{-\left(\frac{m+k+1}{2}\right)} \exp\left\{-\frac{1}{2}\mathrm{tr}\Lambda^{-1}Y^{-1}\right\} \qquad (2.5.12)$$

for k number of variables. When k = 1, the above equation reduces to the inverted chi-squared distribution, as expected.

Incorporating what the data suggest in the form of the likelihood to the assumed prior, and obtaining the posterior via multiplying the prior and likelihood, we derive the posterior distributions:

$$\mu \mid \Sigma, Y \sim N(\mu'_0, (\tau')^{-1}\Sigma) \qquad (2.5.13)$$

and

$$\Sigma \mid Y \sim W^{-1}(m', \Lambda') \qquad (2.5.14)$$

where

$$\begin{aligned} \tau' &= \tau + n \\ m' &= m + n \\ \mu'_0 &= \left(\frac{n}{\tau+n}\right)\bar{y} + \left(\frac{\tau}{\tau+n}\right)\mu_0 \\ \Lambda' &= \left[\Lambda^{-1} + nS + \left(\frac{n}{\tau+n}\right)\left(\bar{y}-\mu_0\right)\left(\bar{y}-\mu_0\right)^T\right]^{-1} \end{aligned} \qquad (2.5.15)$$

These two steps are iterated until convergence in the parameter estimates is reached. Schafer (1997) notes the parallels between the I-step and the E-step and between the P-step and the M-step of the EM algorithm in this aspect.

2.6 Multiple Imputation for Categorical and Mixed Data

Aside from multiple imputation methods under the normality assumption, other approaches have been established to handle missing categorical or binary data and mixed data consisting of both categorical and continuous variables. In Sections 2.4 and 2.5, approaches under the normality assumption were mentioned as a possible manner to impute values for ordinal and binary data. Two other models available for imputation of categorical data include the saturated multinomial model and the loglinear model (Schafer, 1997; Schafer and Olsen, 1998). The multinomial model is based on the probability distribution:

$$P(x \mid \theta) = \frac{n!}{x_1!x_2!...x_D!}\theta_1^{x_1}\theta_2^{x_2}...\theta_D^{x_D} \qquad (2.6.1)$$

where $x = \{x_1, x_2, ..., x_D\}$ and x_d corresponds to the d^{th} cell of a contingency table. Here, we assume that the parameter θ follows a Dirichlet distribution:

$$P(\theta \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_D)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}...\theta_D^{\alpha_D-1} \qquad (2.6.2)$$

The Dirichlet distribution is related to the standard gamma distribution by:

$$P(\nu \mid a) = \frac{1}{\Gamma(\alpha)}\nu^{a-1}e^{-\nu} \qquad (2.6.3)$$

in that we can express a parameter θ_d :

$$\theta_d = \frac{\nu_d}{\sum_{d'=1}^{D} \nu_{d'}}, d = 1, 2, ..., D \qquad (2.6.4)$$

where $\theta = \{\theta_1, \theta_{2,} ..., \theta_D\}$ has a Dirichlet distribution with parameter $\alpha = \{\alpha_1, \alpha_{2,} ..., \alpha_D\}$.

As under the normality assumption, we can implement the EM and DA algorithms in the multiple imputation techniques for discrete and mixed data. Further proceeding with the imputation method with respect to a Bayesian perspective, we can choose a prior depending on the nature of the data at hand. For example, setting the hyperparameter of the Dirichlet distribution α to a constant *c* tends to determine the type of prior and how much information the prior with provide. *c* = 0, 1, or 1/2, for instance, result in the improper, uniform, and Jeffrey's prior, respectively. Likewise, a constant *c* > 1 leads to a flattening prior, adding a constant of ε = *c* – 1 > 0 in each cell of a contingency table, which in turn leads to a more uniform distribution of θ. Such a prior is often recommended when working with sparse data, where a substantial number of cells could have zero counts.

Schafer (1997) also reviews the loglinear model:

$$\eta = M\lambda \qquad (2.6.5)$$

and

$$\eta_d = \log \theta_d, d = 1, 2, ..., D \qquad (2.6.6)$$

i.e., the logarithm of the parameter given in the multinomial distribution (2.6.1). Here λ is the *r* x 1 parameter vector and *M* is the *D* x *r* design matrix determining the constraints for the model. An attractive feature of the loglinear model over the saturated multinomial model is that we can eliminate terms that may not prove necessary. For example, given a model with three categorical variables, we can remove the three-way interaction or the three-way interaction and any two-way interactions deemed insignificant in a hierarchical model. For estimating the parameters of this model, Schafer (1997) discusses the replacing the M-step in the EM algorithm by the CM-step (i.e., conditional maximization), leading to the E-CM algorithm, and iterative proportional fitting (IPF). The CM-step of the E-CM algorithm takes into account constraints imposed on the parameters of the restricted model. Additionally, iterative proportional fitting (IPF) is an iterative procedure by which the parameter of θ is proportionally adjusted to satisfy a series of moment equations until the estimate of θ is stabilized, given that the initial values used to set θ satisfy the loglinear model constraints. Bayesian iterative proportional fitting is introduced as an avenue for simulating random draws from a constrained Dirichlet prior, i.e., a prior employed to satisfy constraints imposed by the loglinear model which can be implemented in the imputation technique. Namely, this approach is combined with data augmentation to form the hybrid data augmentation-Bayesian iterative proportional fitting (DABIF) approach.

We conclude this section by introducing the general location model (Schafer, 1997), which allows imputation of data including both continuous and categorical variables, where the categorical variables could be completely missing, the categorical variables and a subset of the continuous variables are missing, or a subset of categorical and a subset of continuous variables are missing. Taking $W_1,...,W_p$ as the categorical variables and $Z_1,...,Z_q$ as the continuous variables in a data set of dimension $n \times (p+q)$, we define the general location model using:

$$w \,|\, \pi \sim M(n, \pi) \tag{2.6.7}$$

and

$$z_i \,|\, u_i = E_d, \mu_d, \Sigma \sim N(\mu_d, \Sigma) \tag{2.6.8}$$

where $\pi = \{\pi_w : w \in W\}$, E_d is a *D*-vector with 1 at position *d* and 0 elsewhere, μ_d is a *q*-means vector for cell *d*, and Σ is a $q \times q$ covariance matrix. These quantities are computed for *d* = (1,…,*D*) cells in the contingency table corresponding to *p* categorical variables.

The estimation of parameters for such an unrestricted model, i.e., a model containing all possible parameters, involves both multinomial and normal distributions. With data including cells having zero counts in the contingency table, this unrestricted model may not be practical unless the sample size is large, however. In such circumstances, a general location model can be constructed incorporating the loglinear and normal models, such that restrictions are imposed upon parameters of the loglinear model and determine cell probabilities and the distribution of the continuous variables $Z_1,...,Z_q$ conditional on the categorical variables $W_1,...,W_p$ via:

$$Z = U\mu + \varepsilon \qquad (2.6.9)$$

where U is an $n \times D$ matrix with indicator variables for cell location $1, 2, ..., D$ and μ is a $D \times q$ matrix of means. Namely, we can restrict μ using

$$\mu = A\beta \qquad (2.6.10)$$

for a constant $D \times r$ matrix A and some β. Data having zero count cells can now be estimable depending on the rank of the matrix *A*. The likelihood for this restricted model can be computed with the iterative proportional fitting approach discussed earlier.

With the unrestricted general location model, data is imputed using the EM and DA algorithms, where the multinomial and normal distributions are both considered in the P-step of the data augmentation algorithm, involving updating parameters of the multinomial distribution which follow the Dirichlet distribution, and updating mean and covariance parameters of the normal distribution which follow the multivariate normal and inverse Wishart distributions, respectively. With the restricted general location model, the multiple imputation method is similar to that for imputation with the loglinear model, involving the E-CM and the DABIF hybrid algorithms.

2.7 Multiple Imputation Assessment

Several measures are available for assessing the validity of results obtained from imputed data. These measures include standardized bias (SB), percentage bias (PB), coverage rate (CR), root mean-square-error (RMSE), and average width of the confidence interval (AW) (Demirtas and Hedeker, 2007; Demirtas *et al.*, 2008; Demirtas and Hedeker, 2008) SB and PB, given by:

$$100 \times \left|\frac{E(\hat{\theta}-\theta)}{SE(\hat{\theta})}\right| \tag{2.7.1}$$

and

$$100 \times \left|\frac{E(\hat{\theta}-\theta)}{\theta}\right| \tag{2.7.2}$$

respectively, are used to examine the effect of bias on our estimate in either direction. Any SB > 40% - 50% or PB > 5% indicates that the relative magnitude of the absolute value in the bias measures to the estimate can have an adverse effect on the inferences of our estimate. The coverage rate (CR) is the percentage of times the true parameter is encompassed by the confidence interval for the parameter estimate. Collins *et al*. (2001) indicate that coverage rates below 90% imply poor coverage. The root mean square error, RMSE, defined by:

$$\sqrt{E_{\theta}(\hat{\theta}-\theta)^2} \tag{2.7.3}$$

evaluates both variance and bias and provides arguably the best assessment for combined precision and accuracy. Lastly, the average confidence interval width (AW) corresponds to the average difference between the lower and upper confidence limits across each set containing $m > 1$ multiply imputed data. Ideal accuracy and efficiency scenarios are characterized by high CR and narrow AW (Collins *et al.*, 2001; Demirtas and Hedeker, 2007), along with small bias and RMSE.

2.8   Multiple Imputation Software

In proceeding with the different imputation approaches, there are several software packages available, depending on which procedure is necessary (Horton and Kleinman, 2007; Schafer and Graham, 2002; Schafer and Olsen, 1998).   For instance, PROC MI and PROC MIANALYZE in SAS are used in implementing the MCMC approach for Gaussian, parsimonious Markov, regression, logistic, polytomous, and discriminant models. ICE in Stata, MICE in R and S-plus, and IVEware (Imputation and Variance Software), run in SAS or independently, have been devoted to carrying out the chained equation approach (van Buuren *et al*., 1999). Horton and Kleinman (2007) also describe packages such as the NORM, CAT, MIX, and PAN packages in R associated with the joint modeling approach discussed in Section 2.4 (Schafer, 1997; Schafer and Olsen, 1998) to handle missing multivariate normal, categorical, mixed, and longitudinal data, respectively. We use the NORM package in our code, given in the appendix, in implementing our method. Statistical software packages as SOLAS and SPSS also include programs for handling missing data via multiple imputation (Horton and Kleinman, 2007).

3. LURIE-GOLDBERG ALGORITHM

3.1 Lurie-Goldberg Introduction

Thus far, we have discussed multiple imputation, particularly, imputation under normality assumptions, as a computationally favorable approach for handling missing data. Multiple imputation under the normality assumption is restrictive; incorporating these methods, nevertheless, with the Lurie and Goldberg (1998) algorithm can allow us to impute data following any distribution. The Lurie and Goldberg (1998) algorithm involves a technique for generating multivariate random variables using partially specified distributions, meaning that they consider marginal distributions and pairwise correlations. Their method does not require input of the joint distribution with potentially unknown information. Simulation results show that parameter estimates and correlations for data generated through their method closely resemble those for the original data, indicating the benefits of this algorithm. This method incorporates available data in determining relationships between several variables, without collecting more data which may be costly to obtain and allows for generation of data following any continuous, strictly increasing distribution function via a joint normal model.

3.2 Advantages of the Lurie-Goldberg Algorithm Over Other Methods

In their work, Lurie and Goldberg (1998) refer to the PIT, or probability integral transformation, method from Li and Hammond (1975) for generating random variables with non-normal probability distribution functions. Li and Hammond (1975) state that for two known probability distribution functions, say $f_V(v_i)$ and $f_Y(y_i)$, there exists a set of monotone functions, such that:

$$\int_{-\infty}^{v_i} f_{V_i}(v_i)dv_i = \int_{-\infty}^{y_i} f_{Y_i}(y_i)dy_i = F_{Y_i}(y_i) \tag{3.2.1}$$

for $i = 1, \ldots, n$ observations.

Here, $F_y(y_i)$ is the cumulative distribution function. If $f_V(v_i)$ follows the standard normal distribution, then:

$$y_i = F_{Y_i}^{-1}\left[\Phi\left(\frac{v_i}{\sigma_{v_i}}\right)\right] = h_i(v_i) \tag{3.2.2}$$

Where Φ is the standard normal probability distribution function and h_i is the nonlinear transformation. If v_i has unit variance, then (3.2.2) reduces to

$$y_i = F_{Y_i}^{-1}\left[\Phi\left(v_i\right)\right] = h_i(v_i) \tag{3.2.3}$$

where F^{-1} is the inverse cdf for the observed distribution of random variables $y_1, \ldots, y_n$. Li and Hammond (1975) also use probability integral distribution function to derive pairwise correlations between any two variables in the data set. These pairwise relationships make up the entries of a correlation coefficient matrix for the multivariate data. They make note of the requirements for this matrix to be symmetric and positive semidefinite. Given that the pairwise correlation for, say, y_i and y_j, is defined by:

$$\rho_{y_i y_j} = \frac{E\left[y_i y_j\right]}{\sigma_{y_i}\sigma_{y_j}} = \frac{1}{\sigma_{y_i}\sigma_{y_j}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_i(v_i)h_j(v_j)\, f_{v_i v_j}(v_i, v_j)dv_i dv_j, \tag{3.2.4}$$

with $1 \le i \le j \le n$ and employing (3.2.3), they show that the pairwise correlations can be expressed in terms of the probability integral distribution function as:

$$\rho_y = \frac{1}{\sigma_{y_i}\sigma_{y_j}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{Y_i}^{-1}\left[\Phi\left(v_i\right)\right] F_{Y_j}^{-1}\left[\Phi\left(v_j\right)\right] \frac{1}{2\pi\sqrt{1-\rho_v^2}} \exp\left[-\frac{v_i^2 - 2\rho_v v_i v_j + v_j^2}{2(1-\rho_v^2)}\right] dv_i dv_j, \qquad (3.2.5)$$

$$1 \le i \le j \le n$$

where with $\rho_y = \rho_{y_i y_j}$ and $\rho_v = \rho_{v_i v_j}$.

Lurie and Goldberg (1998) note that this probability integral transformation method can be tedious, however, for large numbers of variables. For example, with *k* variables, *k*(*k* -1) probability integral distribution functions would have to be performed. Furthermore, the complex numerical integrations required by Li and Hammond (1975) for each transformation itself may become computationally expensive. Lurie and Goldberg (1998) emphasize that their method eliminates the necessity of numerical integration, thus reducing time and power required for computations.

Lurie and Goldberg (1998) also discuss the advantage of their number generating method over the previous method presented in Vale and Maurelli (1983). In their work, Vale and Maurelli (1983) exploit Fleishman's (1978) polynomials (Demirtas and Hedeker, 2008), such as:

$$Y = a + bX + cX^2 + dX^3 \qquad (3.2.6)$$

where $X \sim N(0, 1)$ is a random variable. The distribution of *Y* then depends on the constants *a, b, c,* and *d*. Bivariate normal variables, as x_1 and x_2 for example, can be drawn and the coefficients (*a, b, c, d*), which can be computed via as set of nonlinear equations, are then used to transform these bivariate normal variables into the desired non-normal data (Vale and Maurelli, 1983). Lurie and Goldberg (1998), however, point out that this polynomial method requires calculation of third and fourth moments, whereas their approach avoids such additional computations. As a secondary point, Lurie and Goldberg (1998) also note that Vale and Maurelli's (1983) method cannot be applied to distributions with bounded support (e.g., Beta).

3.3 Logistics of the Lurie-Goldberg Algorithm

As noted in Section 3.1, the algorithm Lurie and Goldberg (1998) can be implemented in a multiple imputation context to relax restrictive normality assumptions by imposing them not directly on the data but on normally distributed values obtained from transformations of these data. In the Lurie-Goldberg algorithm, the objective of the simulations is to minimize the distance:

$$\underset{\{l_{ij}\}}{Minimize}\_D = \frac{1}{2} \| \mathbf{R}^* - \mathbf{L}\mathbf{L}^T \|^2 \qquad (3.3.1)$$

where **R*** is the correlation matrix with entries containing pairwise correlation estimates from the observed data, **L** is the lower triangular matrix derived from the Cholesky decomposition of the correlation matrix associated with the generated data, and l_{ij} are the elements of **L**. If this correlation is non-positive semidefinite, a probable result when pairwise correlations are calculated separately for data with variables having different missing data patterns, then the Lurie-Goldberg algorithm can be used to generate a positive semidefinite matrix "closest" to the non-positive semidefinite correlation matrix at hand. Lurie and Goldberg (1998) first generate *nk* i.i.d. $N(0,1)$ variables and arrange them in an *n* x *k* matrix, **X**. They then multiply **X** to the transpose of **L**, obtaining:

$$\mathbf{Y} = \mathbf{X}\mathbf{L}^{\mathbf{T}} \qquad (3.3.2)$$

Therefore, $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$ and $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{R})$. Next, they derive the standard normal CDF, or cumulative distribution function, for each entry in **Y**, yielding:

$$\mathbf{U} = \Phi(\mathbf{Y}) \qquad (3.3.3)$$

where Φ represents the standard normal CDF for each entry in **Y**.

Lastly, the authors present the generated data in the matrix **V** with entries:

$$v_{ij} = F_j^{-1}(u_{ij}) \tag{3.3.4}$$

where they use a separate F_j^{-1} function based on the marginal distribution of each *jth* variable in their data. These entries are now in the same scale as their original data. The updated correlation matrix, **R = LL**^{T}, is then computed for the data of the newly generated **V** matrix. Lurie and Goldberg (1998) re-iterate their steps until their root mean square error, RMSE is less than some constant *c*, determined by the desired accuracy. For example, the authors recommend setting *c* = 0.005 if two-digit accuracy is desired. The RMSE is given by:

$$RMSE = \sqrt{4D/[k(k-1)]} \tag{3.3.5}$$

where *D* is the squared norm of the absolute difference between the target and generated correlation matrices and *k* the number of variables in the data. *D* can be expressed as in (3.3.1), involving Cholesky decomposition or can be expressed as:

$$\frac{1}{2}\sum_{i=2}^{k}\sum_{j=1}^{k-1}(r_{ij}^{*} - r_{ij})^{2} \tag{3.3.6}$$

where r_{ij}^{*} and r_{ij} are the elements of *R** and *R*, the target correlation matrix and the correlation matrix associated with the generated data, respectively, and *R** may be positive semidefinite or non-positive semidefinite. Lurie and Goldberg's (1998) simulation results prove promising by the comparable parameter estimates between their original and generated data. We summarize the algorithm presented here as:

$$\mathbf{X} \rightarrow \mathbf{Y} \rightarrow \mathbf{U} \rightarrow \mathbf{V} \tag{3.3.7}$$

4. EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION (eCDF)

The eCDF (empirical cumulative distribution function) values are defined for a real value *x* of a random variable *X* as the proportion of values less than *x*, as given in (4.1.1). These values serve as an avenue for relating a probability to a certain data value without making any specific distributional assumptions. Therefore, we could use eCDF calculations instead of information based on specific marginal distributions to obtain probabilities values involved in transformations and back-transformations of data discussed in Lurie and Goldberg (1998).

4.1 Computing the empirical Cumulative Distribution Function

We have seen how the Lurie and Goldberg (1998) method can be implemented under any distributional assumption. Since the specific distribution usually remains unknown, however, makes consideration of incorporating calculations of eCDF values instead of CDF values based on a specific distribution into our algorithm favorable. We compute the eCDF using:

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \qquad (4.1.1)$$

Incorporation of this quantity is explained in Section 6.1.

4.2 Back-transformation of empirical Cumulative Distribution Function Values

We can also back-transform newly generated eCDF values to values within the range of the original data of interest via the method given in Barton and Schruben (1993). In this approach, we determine an interval, $\left\{F(y_{(i)j}), F(y_{(i+1)j})\right\}$, with two original eCDF values obtained from the observed data encompassing each new eCDF value $u_{C_{(i)j}}$ for each ordered observation $y_{(i)j}$ in variable **Yj**, j = 1, . . . ,*k*. Then, we calculate the difference between the new value and the lower end of the empirical cumulative distribution interval and divide this difference by the length of the interval, $\left\{F(y_{(i)j}), F(y_{(i+1)j})\right\}$. Next, we multiply the outcome from this division by the length of the corresponding interval for the original data values, i.e., $(y_{(i+1)j} - y_{(i)j})$, and add this product to the lower original data value of the corresponding interval such that:

$$F^{-1}\left(u_{C_{(i)j}}\right) = y_{(i)j} + (y_{(i+1)j} - y_{(i)j})\frac{u_{C_{(i)j}} - F(y_{(i)j})}{F(y_{(i+1)j}) - F(y_{(i)j})} \tag{4.2.1}$$

$F^{-1}\left(u_{C_{(i)j}}\right)$ thus maps the eCDF value in question to the scale of the original data.

5. GENERATING BINARY AND MIXED DATA

Here, we discuss generation of binary and mixed data using data following an underlying normal distribution. We present this section as an introduction to imputing binary and mixed data, where data will be imputed under the normality assumption and then transformed into the desired binary values (Section 6.2) or mixed data (Section 6.3) via these described methods.

5.1 Generating Binary Data from Normal Data

Assuming that we have two binary variables, Y_1 and Y_2, we can compute a cross-tabulation of the data, as shown in Table I.

Table I: CROSS-TABULATION OF BINARY VARIABLES Y_1 AND Y_2

Y_1	Y_2	
	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

Then, we can calculate the correlation coefficient, phi, a derivative of the Pearson correlation (Guilford, J., 1936) by:

$$\frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{10}+n_{11})(n_{00}+n_{01})+(n_{00}+n_{10})(n_{01}+n_{11})}} \qquad (5.1.1)$$

where phi can range from:

$$\delta_{jk} \in \left\{ \max\left(-\sqrt{(p_j p_k / q_j q_k)}, -\sqrt{(q_j q_k / p_j p_k)}\right), \min\left(\sqrt{(p_j q_k / q_j p_k)}, \sqrt{(q_j p_k / p_j q_k)}\right)\right\},$$
$$p_j = \Pr(Y_j = 1), q_j = 1 - p_j \tag{5.1.2}$$

We can obtain the tetrachoric correlation ρ_{jk} using:

$$\Phi\left[z(p_j), z(p_k), \rho_{jk}\right] = \delta_{jk}(p_j q_j p_k q_k)^{1/2} + p_j p_k \tag{5.1.3}$$

(Emrich and Piedmonte, 1991; Demirtas and Doganay, 2012) and generate a standard bivariate normal data set, Z, and covariance, $\begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix}$, and introduce the same fraction of missing entries in this data set as found in the original data set. We can then introduce some probabilities involving Y_1 and Y_2 pertaining to quantiles in the bivariate normal data which in turn allow for the preservation of the same proportions observed in the original binary data. We will provide further description of these quantiles will discussed in Section 6.2 and how they help us define binary values from data imputed under the normality assumption.

<u>5.2 Tetrachoric and Polychoric Correlations</u>

5.2.1 Definition of Tetrachoric and Polychoric Correlations

Several works discuss the advantage of the tetrachoric and polychoric correlation over the Pearson correlation when estimating associations between binary and ordinal variables. Here, the tetrachoric correlation coefficient ρ_{jk} given in the previous equation (5.1.3) for Y_j = I($Z_j \leq z(p_j)$) is a special case of the polychoric correlation used with binary data (Emrich and Piedmonte, 1991; Demirtas and Doganay, 2012). The polychoric correlation coefficient can be defined for ordinal data, where:

$$Y_j = \begin{cases} 1, Z_j \leq z(p_{1j}) \\ 2, z(p_{1j}) \leq Z_j \leq z(p_{2j}) \\ 3, z(p_{2j}) \leq Z_j \leq z(p_{3j}) \\ \vdots \\ s, z(p_{(s-1)j}) \leq Z_j \end{cases} \tag{5.2.1}$$

and can be computed via maximum likelihood estimation with acceptable accuracy (Olsson, 1979).

5.2.2 Advantages of Tetrachoric and Polychoric Correlations

A common notion is that obtaining the tetrachoric and polychoric correlations are less biased towards zero than calculating the Pearson correlation directly from underlying normally distributed variables, albeit their estimated standard errors may be slightly larger (Babakus et al., 1987; Butler *et al.*, 1987; Rigdon and Ferguson, 1991). Simulations have also been conducted to examine the performance of these measures. Work by Babakus et al. (1987) show that the polychoric correlation produces better results in terms of precision and accuracy than computing the Pearson and Spearman correlations and Kendall's tau coefficient directly with ordinal data. Rigdon and Ferguson (1991) evaluate the performance of the polychoric correlation in combination with functions as unweighted least squares, weighted least squares, generalized least squares, and diagonally weighted least squares used in model fitting. They note weighted least squares as optimal in combination with the polychoric correlation for covariate estimation in models applied to ordinal data, leading to decreased bias in parameter estimation. Some drawbacks of this approach, however, involve a slightly higher rejection of the correct model in certain scenarios. Questionable estimation via methods combined with the polychoric correlation could also arise when applied to small sample sizes or skewed data.

Nevertheless, tetrachoric and polychoric correlations can still be favorable over other correlation measures and the odds ratio in terms of estimation. For instance, the tetrachoric may be more practical to employ when examining associations of several binary variables (Le Cessie and Van Houwelinden, 1994; Qu et al., 1995), where odds ratios, albeit not restricted to the (-1, 1) range and therefore easier to interpret, cannot be feasibly obtained due to difficulty associated with calculation of the full likelihood. Tetrachoric and polychoric correlations also can be applied to repeated measures data to measure within-subject variation with respect to between-subject variation, thus serving as an equivalent to the intraclass correlation coefficient. Qu et al. (1995) exemplify this use of the coefficient to estimate the correlation between the face and arms in a study involving a double-blinded randomized clinical trial examining the effects of TEC medication on premature skin aging caused by ultraviolet radiation. They further show the validity of the polychoric correlation in measuring within-subject correlations in GEE models via simulation studies. Thus, tetrachoric and polychoric correlations have been proven as a useful association measure in several cases involving binary and ordinal data.

5.2.3 Software for Computing Tetrachoric and Polychoric Correlations

Several software packages include modules to compute tetrachoric and polychoric correlations. For example, in SAS, the PLCORR option is available for this computation in the TABLES statement of the FREQ procedure. Alternatively, a %POLYCHOR macro can be downloaded from the SAS support website at http://support.sas.com/kb/25/010.html, based on the maximum likelihood approach given in Olsson (1979) and Drasgow (1986). An R 'polycor' library includes a 'polychor' function written by John Fox for calculating tetrachoric and polychoric correlations also via maximum likelihood estimation. Furthermore, the 'phi2poly' function in the R 'psych' (Revelle, 2011) employs the 'polycor' library and can be used compute tetrachoric correlations from the phi correlation coefficient for binary data. We use this latter 'phi2poly' function in our method for imputing binary data. Likewise a 'r_tetra' macro has been written by Dirk Enzmann computing the tetrachoric correlation can be executed in SPSS. STATA also includes a tetrachoric command and a polychoric command by Stas Kolenikov for computing these correlations (Uebersax, 2011).

5.3 Point-biserial Correlation

The point-biserial correlation allows investigators to measure the association between a continuous and binary variable, where the binary nature of the latter variable can be inherent, as in the case of sex or smoking status (Tate, 1954; Demirtas and Doganay, 2012) or can be derived from the dichotomization or a continuous variable. This latter approach is sometimes favorable in clinical, psychological, or ecomomic settings where they provide easier interpretation of the data as in cases of defining obesity from BMI values (Demirtas and Doganay, 2012) or categorizing psychological data collected on a continuous scale in order to predict juvenile delinquency (Farrington and Loeber, 2000). The point-biserial correlation can be defined as:

$$\delta_{Y_1Y_{2D}} = \frac{\mu_{1\{Y_{2D}=1\}} - \mu_{1\{Y_{2D}=0\}}}{\sigma_1} \tag{5.3.1}$$

where $\mu_{1\{Y_{2D}=1\}}$ is the mean of the continuous variable Y_1 where the binary variable $Y_{2D} = 1$, $\mu_{1\{Y_{2D}=0\}}$ is the mean of the continuous variable Y_1 where the binary variable $Y_{2D} = 0$, and σ_1 the is variance of Y_1. Furthermore, if the binary variable was derived via dichotomization from an inherently continuous variable, we can define the relationship between the correlation of two normally distributed variables and of two originally normally distributed variables, with one variable dichotomized as:

$$\begin{gathered}\delta_{Y_1Y_{2D}} = \left(\frac{h}{\sqrt{p(1-p)}}\right)\rho_{Y_1Y_2}, \\ p = \Pr(Y_{2D} = 1)\end{gathered} \tag{5.3.2}$$

where h is the ordinate of the normal curve at some point X such that:

$$h = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{\frac{-(X-\mu)^2}{2\sigma^2}\right\} \tag{5.3.3}$$

involving the N(μ, σ) distribution. In Section 6.3, we show how an alternative expression of equation (5.3.2) relates to a multivariate normal data set that can then be imputed under the normality assumption. Since the point-biserial correlation is a product-moment coefficient, it can be computed between any continuous and binary variable using the command for calculating the Pearson correlation in any software package, as the 'cor' function in R, PROC CORR in SAS, the CORRELATION option in SPSS, and the 'corr' option in STATA.

5.4 Correlation Bounds

Demirtas and Hedeker (2011) discuss correlation bounds spanning a range narrower than (-1, 1) as seen with the tetrachoric and point-biserial correlations in equations (5.1.2) and (5.3.2), respectively. In cases where these bounds are not easily computed in closed form, the authors recommend generating data with a large number of observations of the intended distribution in the same and opposing directions and then calculate the correlations, giving us the maximum correlation and anti-correlation, respectively. Their simulation results support the validity of this method. The authors encourage programmers to use this method to find correlation bounds that are otherwise difficult to obtain in closed form. Furthermore, their method could be used to compute correlation matrices with pairwise elements derived from different sources (Lurie and Goldberg, 1998). Demirtas and Hedeker (2011) conclude by emphasizing the importance of checking correlation bounds before proceeding with any simulation study.

6. PROPOSED SEMI-PARAMETRIC METHODS FOR IMPUTING DATA

We now present our approaches for imputing continuous, binary, and mixed data based on methods described in the previous sections. We first apply the concept of multiple imputation under the normality assumption in Section 2.5, the Lurie-Goldberg algorithm in Section 3 and eCDF calculations in Section 4 to impute continuous data (Section 6.1). Imputing binary data (Section 6.2) also employs multiple imputation under the normality assumption, as well as concepts of generating binary data from normal data where the tetrachoric correlation is a measure of association between variables (Sections 5.1 and 5.2). We conclude with imputing mixed data based on multiple imputation under the normality assumption and generating mixed data from normally distributed values and involving the point-biserial correlation (Section 5.3).

6.1 Imputing Continuous Data

Using our notation of Y_{com} for complete data, Y_{obs} for observed data, and Y_{mis} for missing data, we first introduce our proposal for imputing continuous data by incorporating aspects of multiple imputation under normality assumptions, the Lurie-Goldberg (1998) algorithm, and eCDF calculations. We first create a matrix, $\mathbf{U}_{com}$, containing the eCDF values for the observed data and missing entries for corresponding missing values in the data. Next, we use the inverse function F^{-1} separately for each variable, where F is the marginal N(0,1) distribution function to obtain:

$$\mathbf{Y}^{*}_{com} = F^{-1}(\mathbf{U}_{com}) \tag{6.1.1}$$

The covariance matrix of $\mathbf{Y}^{*}_{com}$ is then the correlation matrix for Y_{com}. The multiple imputation method under the normality assumption is then applied, as in Schafer (1997), leading to:

$$\mathbf{Y}_{com}^{*imp} \sim N(\mu^{imp}, \Sigma^{imp}) \tag{6.1.2}$$

where $\left(\mu^{imp}\right)' = \mathbf{0}$ and Σ^{imp} are the mean-vector and variance-covariance matrix for the imputed data obtained via the EM and DA algorithms, respectively. $\mathbf{Y}_{com}^{*imp}$ is then the matrix containing the imputed data as well as the observed transformed data. We back-transform this data to obtain:

$$\mathbf{U}_{com}^{imp} = F(\mathbf{Y}_{com}^{*imp}) \tag{6.1.3}$$

with *F* being the cumulative distribution function based on the normal distribution with updated parameters. We finally obtain our originally observed values and map the imputed values to the range of the original data using the method described in Section 4.2 from Barton and Schruben (1993). Defining a matrix $\mathbf{Y}_{com}^{imp}$ containing these values, we summarize our method with the following diagram:

$$\begin{aligned}
&Y_{com}\\
&1. \rightarrow U_{com} = eCDF(Y_{com})\\
&2. \rightarrow Y_{com}^{*} = \Phi^{-1}(U_{com})\\
&3. \rightarrow Y_{com}^{*imp}\\
&4. \rightarrow U_{com}^{imp} = F(Y_{com}^{*imp})\\
&5. \rightarrow Y_{com}^{imp}
\end{aligned} \tag{6.1.4}$$

We re-iterate the steps (3) to (5) until the absolute difference between our generated correlations and the target correlation is less than the product of the target correlation multiplied by some constant c_{jk}.

i.e., $$\left|\rho_{jk} - \rho_{jk}^{imp}\right| < c_{jk}\rho_{jk} \qquad (6.1.5)$$

for each pairwise correlation between variables Y_j and Y_k, $j = 1, \ldots, p - 1$, $k = 2, \ldots p$ in a data set with p variables, and ρ_{jk} and ρ_{jk}^{imp} are the pairwise correlations between variables Y_j and Y_k, respectively. The recommended range for c_{jk} is (0.01, 0.05) and the choice of this constant depends on minimizing the bias and maximizing the coverage rate associated with each pairwise correlation coefficient.

Our algorithm can be compared to that found in Lurie and Goldberg (1998) in that a key component of it involves transformations of normally distributed variables. In the case of Lurie and Goldberg (1998), however, all initial values are randomly drawn from a N(0,1) distribution, whereas normally distributed values are obtained via the inverse standard normal distribution function applied to eCDF values of the original data in our case. Additionally, only imputed values used to fill in missing entries involve random draws in our case. Furthermore, the Lurie and Goldberg (1998) algorithm employs the inverse functions of specific marginal distributions to transform their normally distributed data to data with variables following the desired distributions, but we map the CDF values based on the normal distribution onto the scale of the original data using the inverse function for eCDF values as described in Barton and Shruben (1993) allowing for nonparametric back-transformations of the data. Similarities and differences between the Lurie and Goldberg (1998) and our imputation method continuous data are presented in Table II.

Table II: SIMILARITIES AND DIFFERENCES INVOLVING METHODS FOR CONTINUOUS DATA

Steps	Lurie-Goldberg (1998) algorithm	LGMI (2011) algorithm
1	Normally distributed data are generated via random draws from N(0,1) distribution and specified pairwise correlations are induced via Cholesky decomposition.	Normally distributed values are obtained via calculating eCDF values for original data and then the inverse distribution function based on the N(0,1) distribution is applied to these eCDF values; pairwise correlations from the original data are preserved in this case.
2		Multiple imputation via joint modeling under the normality assumption is applied to the multivariate normally distributed data.
3	The inverse functions of specified marginal distributions are employed to map all generated values onto the scale of the final data set created.	The inverse functions based on marginal eCDF values are employed to map only imputed values onto the scale of the final data set created.

6.2 Imputing Binary Data

With the binary data generation techniques described in Section 5.1, we proceed with computing quantiles that will allow us to create binary data from the imputed normal data. First, we assume that we have two binary variables, Y_1 and Y_2, where some proportion of Y_2, $P(R_2 = 0)$, is missing, we can compute a cross-tabulation of the data, as given in Table III.

Table III: CROSS-TABULATION WITH Y_1 ANDY_2, WHERE Y_2 INCLUDES MISSING VALUES

Y_1	Y_2		
	0	1	?
0	n_{00}	n_{01}	$n_{0?}$
1	n_{10}	n_{11}	$n_{1?}$

We then calculate

$$\begin{aligned} P(Y_1=1,Y_2=1) &= P(Y_1=1,Y_2=1,R_2=1) = P(Y_2=1 \mid Y_1=1,R_2=1)P(Y_1=1,R_2=1) \\ &= P(Y_2=1 \mid Y_1=1,R_2=1)P(Y_1=1 \mid R_2=1)P(R_2=1) \end{aligned} \tag{6.2.1}$$

and obtain corresponding quantiles given by

$$\begin{aligned} q_{11} &= Q_{P(Y_1=1|R_2=1)}(Z_1) \\ q_{21} &= Q_{P(Y_2=1|Y_1=1,R_2=1)}(Z_2); q^*_{21} = Q_{P(Y_2=1|Y_1=0,R_2=1)}(Z_2) \end{aligned} \tag{6.2.2}$$

With these quantiles, we can compute proportions based on the generated bivariate normal data with variables Z_1 and Z_2. The number of observed entries should then be the same as the cell counts given in Table II.

i.e.,

$$\sum I(Y_1 = 1, Y_2 = 1 \mid R_2 = 1) = \sum I(Z_1 < q_{11}, Z_2 < q_{21} \mid R_2 = 1); \sum I(Y_1 = 1, Y_2 = 0 \mid R_2 = 1) = \sum I(Z_1 < q_{11}, Z_2 > q_{21} \mid R_2 = 1)$$
$$\sum I(Y_1 = 0, Y_2 = 1 \mid R_2 = 1) = \sum I(Z_1 > q_{11}, Z_2 < q^*_{21} \mid R_2 = 1); \sum I(Y_1 = 0, Y_2 = 0 \mid R_2 = 1) = \sum I(Z_1 > q_{11}, Z_2 > q^*_{21} \mid R_2 = 1) \quad (6.2.3)$$

where R_2 = 1, 0 for Z_2 observed and missing, respectively.

After obtaining quantiles determined to give us correct counts of the original correlated binary data, we proceed with imputing the normal data by applying the joint modeling approach discussed in Schafer (1997) to our bivariate normal data, Z, and obtain Z_{imp}. We then apply the previously determined quantiles based on the normal values corresponding to observed entries to obtain binary outcomes for the imputed values. Namely, we use quantiles conditional on Y_1 =1 to obtain outcomes for imputed Z values corresponding to entries where Y_1 =1 and quantiles conditional on Y_1 = 0 for imputed values corresponding to entries where Y_1 = 0.

This procedure can also be extended to bivariate data where both Y_1 and Y_2 have missing entries. Note that from the joint probability for Y_1 = 1 and Y_2 = 1 defined earlier, we can further define:

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1) &= P(Y_1 = 1, Y_2 = 1, R_1 = 1, R_2 = 1) = P(Y_2 = 1 \mid Y_1 = 1, R_1 = 1, R_2 = 1) P(Y_1 = 1, R_1 = 1, R_2 = 1) \\ &= P(Y_2 = 1 \mid Y_1 = 1, R_1 = 1, R_2 = 1) P(Y_1 = 1 \mid R_1 = 1, R_2 = 1) P(R_1 = 1, R_2 = 1) \\ &= P(Y_1 = 1 \mid Y_2 = 1, R_1 = 1, R_2 = 1) P(Y_2 = 1 \mid R_1 = 1, R_2 = 1) P(R_1 = 1, R_2 = 1) \end{aligned} \quad (6.2.4)$$

We can thus obtain binary outcomes from imputed Z_1 and Z_2 values for variables Y_1 and Y_2 using quantiles based on:

$$\begin{aligned} q_{11} &= Q_{P(Y_1 = 1 | Y_2 = 1, R_1 = 1, R_2 = 1)}(Z_1); q^*_{11} = Q_{P(Y_1 = 1 | Y_2 = 0, R_1 = 1, R_2 = 1)}(Z_1) \\ q_{21} &= Q_{P(Y_2 = 1 | Y_1 = 1, R_1 = 1, R_2 = 1)}(Z_2); q^*_{21} = Q_{P(Y_2 = 1 | Y_1 = 0, R_1 = 1, R_2 = 1)}(Z_2) \end{aligned} \quad (6.2.5)$$

Again, quantiles are based on data corresponding to entries with both variables observed, i.e., R_1 = 1 and R_2 = 1.

We further extend our method to the multivariate case by basing our quantiles on probabilities:

$$\begin{aligned}&\Pr(Y_k = y_k \mid Y_1 = y_1, Y_2 = y_2, \ldots, Y_{k-1} = y_{k-1}, R_1, R_2, \ldots R_{K-1}, R_K),\\ &y_k = 0,1; k = 1,\ldots,K\end{aligned} \tag{6.2.6}$$

where y_k = 0,1, k = 1, . . . $K \geq 3$ and K is the number of variables in our data set.

The conditional probabilities for all K variables can be derived from the joint probability:

$$\begin{aligned}&\Pr(Y_1 = y_1, Y_2 = y_2, \ldots Y_{K-1} = y_{K-1}, Y_K = y_K, R_1, R_2, \ldots R_{K-1}, R_K) = \Pr(Y_1 = y_1 \mid R_1, R_2, \ldots R_{K-1}, R_K) *\\ &\Pr(Y_2 = y_2 \mid Y_1 = y_1, R_1, R_2, \ldots R_{K-1}, R_K) * \Pr(Y_k = y_k \mid Y_1 = y_1, Y_2 = y_2, \ldots Y_{k-1} = y_{k-1}, R_1, R_2, \ldots R_{K-1}, R_K),\\ &y_k = 0,1; k = 1,\ldots,K\end{aligned} \tag{6.2.7}$$

Thus, quantiles can be obtained via:

$$q_k = Q_{\Pr(Y_k = 1 | Y_1, Y_2, \ldots, Y_{k-1}, R_1, R_2, \ldots, R_k)}(Z_k) \tag{6.2.8}$$

leading to:

$$\sum I(Y_k = 1 \mid Y_1, Y_2, \ldots, Y_{k-1}, R_1, R_2, \ldots, R_k) = \sum I(Z_k < q_k \mid R_1, R_2, \ldots, R_k) \tag{6.2.9}$$

for all combinations with $Y_1, \ldots, Y_k \in \{0,1\}$.

Here, Z_1, . . . , Z_K comprise the normally distributed variables of a data set, Z, with mean **0**, and covariance Σ, where the elements are pairwise tetrachoric correlations derived from the pairwise phi correlations via Equation (5.1.3). We then impute data pertaining to the multivariate normal data set, Z, and again dichotomize the newly imputed data for each variable k, k = 1,...,K via the quantiles obtained from Equations (6.2.8) and (6.2.9).

After computing pairwise correlations for the newly imputed binary data, we check if the updated phi matrix containing these pairwise elements is positive definite and if the matrix is fairly close to the original phi matrix, i.e., if for each element,

$$\left|\delta_{jk}^{imp} - \delta_{jk}\right| < c_{jk} \tag{6.2.10}$$

for some constant c_{jk} chosen to minimize standardized bias and maximize coverage rates.

If the new phi matrix is non-positive definite, then we derive the 'nearest' positive definite phi matrix and compare the elements of this matrix to those of the original phi matrix. This technique is summarized in the following diagram in (6.2.11), where steps (2) – (3) are re-iterated until the convergence criteria in (6.2.10) for all pairwise correlations are met.

$$\begin{aligned} &Y \\ &1. \rightarrow Z \\ &2. \rightarrow Z^{imp} \\ &3. \rightarrow Y^{imp} \end{aligned} \tag{6.2.11}$$

As in Emrich and Piedmonte (1991) , we generate multivariate normally distributed data using tetrachoric pairwise correlations and dichotomize the normally distributed variables based on quantiles associated with previously obtained probabilities. The tetrachoric corrlation is a special case of the Pearson correlation relating two normally distributed variables underlying two binary variables. Obtaining the tetrachoric correlations using pairwise phi correlations from the corresponding binary variables is preferable to obtaining the Pearson correlation directly from the generated normally distributed data underlying the binary variables in terms of precision and accuracy and lower bias when measuring associations of the variables, as discussed in Section 5.2.2. Unlike Emrich and Piedmonte (1991), we also introduce missing values in the multivariate normally distributed variables, where the amount of missingness in each variable is equal to the amount of missingness in the corresponding original binary variable. We then obtain quantiles using the variables with normally distributed data including missing entries and only dichotomize imputed data which are then used to fill in missing values in the original binary data. We summarize comparisons between the two methods in Table IV below.

Table IV: SIMILARITIES AND DIFFERENCES INVOLVING METHODS FOR BINARY DATA

Steps	Emrich-Piedmonte (1991) algorithm	MI for binary data (2011) algorithm
1	Multivariate normally distributed data are generated using tetrachoric pairwise correlations.	Multivariate normally distributed data are generated using tetrachoric pairwise correlations.
2		Missing values are introduced in multivariate normally distributed data; the percentage of missing values is equal to the percentage of missing values in the original binary data.
3	Quantiles associated with the multivariate normally distributed data are obtained based on probabilities involving the desired data.	Quantiles associated with the multivariate normally distributed data after missing values are introduced are obtained based on probabilities involving the original data.
4		Multiple imputation via joint modeling under the normality assumption is applied to the multivariate normally distributed data.
5	Generated data are dichotomized with quantiles obtained in Step 3.	Imputed data are dichotomized with quantiles obtained in Step 3.

6.3 Imputing Mixed Data

We could combine principles for imputing continuous data and binary data in order to impute mixed data. We introduce this approach by starting with a bivariate example, where one variable, Y_1, is continuous and the other variable, Y_2 is binary. Y_1 can be transformed to a normally distributed variable via the eCDF approach given in the LGMI algorithm. With Y_2, we can first generate a random normal variable Z_2 and then re-arrange the values of this variable such that entries corresponding to entries in the original binary variable $Y_2 = 0$ are less than the quantile of the generated data based on the proportion of zeroes in Y_2, defined by probability $\Pr(Y_2 = 0)$ and values corresponding $Y_2 = 1$ are greater than this quantile, i.e, $Q_z = q_{\Pr(Y_2=0)}(z_2)$. We then impose missing values in the normally distributed variable in the same positions of missing entries Y_2. The correlation associated with this new data set is the point-biserial correlation given as:

$$\begin{aligned} &\rho_{Y_1Y_2} = \delta_{Y_1Y_{2D}}\left(\frac{\sqrt{p(1-p)}}{h}\right), \\ &p = \Pr(Y_{2D} = 1) \end{aligned} \qquad (6.3.1)$$

which is an alternative expression of equation (5.3.2).

We then impute this new normally distributed data set and transform them onto the original scale. Y_1 is back-transformed via the Barton and Schruben (1993) method and Y_2 is transformed such that imputed values less than the previously described quantiles are coded as 0 and 1, otherwise. We then compute the point-biserial correlation, $\delta_{Y_1Y_{2D}}$, of the imputed data set and compare this value to the original correlation.

Next, we extend our method for imputing mixed data to multivariate data sets with $k \geq 3$ variables, where $p \geq 2$ variables are binary. Here, we again transform the continuous data via the eCDF approach given in the LGMI method. With binary data, we first generate a multivariate normal data set associated with a tetrachoric correlation matrix derived from the pairwise phi correlation coefficients and delete entries corresponding to missing entries in the original binary data. We then combine these variables with the variables related to the original continuous data. After imputing the data, we back-transform the continuous variables via the Barton and Schruben (1993) method and transform the imputed values in the original binary variables via quantiles. These quantiles are based on the conditional probabilities defined as:

$$\Pr(Y_r \mid Y_1, ..., Y_{r-1}, Y_{r+1}, ..., Y_p) \qquad (6.3.2)$$

for r = 1, . . . p binary variables.

If we define:

$$\begin{aligned} p_r &= \Pr(Y_r = 1 \mid Y_1, ..., Y_{r-1}, Y_{r+1}, ..., Y_p), \\ p_r^{'} &= 1 - \Pr(Y_r = 1 \mid Y_1, ..., Y_{r-1}, Y_{r+1}, ..., Y_p) = \Pr(Y_r = 0 \mid Y_1, ..., Y_{r-1}, Y_{r+1}, ..., Y_p) \end{aligned} \qquad (6.3.3)$$

then we can create binary values in the variable Y_r from the imputed values of Z_r by applying (6.3.4).

$$I(Z_r < Q_{p_r^{'}}(Z_r)) \qquad (6.3.4)$$

We summarize the steps for imputing multivariate mixed data as the diagram in equation (6.3.5).

$$\begin{array}{ll} Y_{cont} & Y_{cont}^{*imp} \\ 1. \to U_{con} & 6. \to U_{cont}^{imp} \\ 2. \to Y_{cont}^{*} & 7. \to Y_{cont}^{imp} \\ & \\ Y_{bin} & Y_{bin}^{*imp} \\ 3. \to Y_{bin}^{*} & 8. \to Y_{bin}^{imp} \\ & \\ 4. Y^{*} = (Y_{cont}^{*}, Y_{bin}^{*}) & 9. Y^{imp} = (Y_{cont}^{imp}, Y_{bin}^{imp}) \\ 5. \to Y^{*imp} & \end{array} \qquad (6.3.5)$$

We re-iterate steps (4) to (9) until the convergence criteria with all pairwise correlations where the criteria are given in equation (6.3.6).

$$\left| \delta_{jk} - \delta_{jk}^{imp} \right| < c_{jk} \qquad (6.3.6)$$

where, for variables Y_j and Y_k, δ_{jk} is the Pearson correlation when Y_j and Y_k are continuous, the phi coefficient when Y_j and Y_k are binary, and the point-biserial correlation when Y_j is continuous and Y_k is binary, and c_{jk} is some constant chosen to optimize estimation of the correlation from the imputed data.

As in Demirtas and Doganay (2012), we generate multivariate normally distributed data using tetrachoric pairwise correlations corresponding to the binary variables in our data set. Instead of generating multivariate normally distributed values corresponding to continuous variables, however, we only map the existent values from the continuous variables to normally distributed data using eCDF computations and the inverse function of the N(0,1) distribution. Additionally, we note that the continuous data is assumed to follow a normal distribution in Demirtas and Doganay (2012), whereas our imputation method assumes that the continuous data can follow any distribution. Furthermore, the same amount of missingness is introduced in the normally distributed variables corresponding to the original binary variables as found in these original variables. Quantiles associated with probabilities from the original data are obtained from these normally distributed variables with missing values. Multiple imputation under the normality assumption is then applied and back-transformation of variables in the data designated as continuous involve the CDF values of imputed data based on the normal distribution and the Barton and Schruben (1993) method and variables in the data designated as binary are dichotomized by the calculated quantiles. These final steps differ from Demirtas and Doganay (2012) in that only imputed and not all generated values are back-transformed and again variables designed as continuous can follow any distribution and not only the normal distribution. These comparisons are summarized in Table V.

Table V: SIMILARITIES AND DIFFERENCES INVOLVING METHODS FOR MIXED DATA

Steps	Demirtas-Doganay (2012) algorithm	MI for mixed data (2011) algorithm
1	Pairwise phi correlations between binary variables, pairwise point-biserial correlations between binary and normally distributed variables, and pairwise Pearson correlations between normally distributed variables are computed.	Continuous and binary variables are separated. Continuous variables are mapped to normally distributed values via eCDF computations and the inverse function of the N(0,1) distribution.
2	Multivariate normally distributed data are generated using tetrachoric correlations associated with phi correlations, biserial correlations associated with point-biserial correlation, and Pearson correlations. (Note: phi, tetrachoric, point-biserial, and biserial correlations are special cases of the Pearson correlation).	Multivariate normally distributed data are only generated for binary variables based on tetrachoric correlations. An amount of missingness is introduced in these data equal to the amount of missingness in the original binary variables.
3	Quantiles associated with the multivariate normally distributed data are obtained based on probabilities for the binary variables.	Quantiles associated with the multivariate normally distributed data after missing values are introduced are obtained based on probabilities for the original binary variables,
4		The multivariate normally distributed data associated with both continuous and binary variables are combined and multiple imputation via joint modeling under the normality assumption is applied to the multivariate normally distributed data.
5	Generated normally distributed data of variables designated as binary are dichotomized by the obtained quantiles.	Imputed normally distributed data of variables designated as binary are dichotomized by the obtained quantiles.
6	Generated normally distributed data of variables designated as following the normal distribution are back-transformed via reverse centering and scaling.	Imputed normally distributed data of variables designated as continuous are back-transformed by obtaining their CDF values based on the normal distribution and mapping these CDF values onto the range of the original continuous data via the Barton and Schruben (1993) method.
7	Normally distributed and binary variables are combined.	Continuous and binary variables are combined.

7. SIMULATIONS WITH GENERATED DATA

7.1 Bivariate Continuous Data

In our first sets of simulations, we generated bivariate data under the assumption of the N(0,1), t_3, and Gamma(1,1) distributions with 500 entries and imposed 50% missingness in the second variable via an MCAR mechanism. Both variables in each bivariate data set followed the same distribution. Under the MCAR mechanism, 50% of entries were deleted from the second variable, Y_2. These data sets were also associated with Pearson correlations ranging from approximately -0.8 to 0.8.

Here, true Pearson correlations and means were obtained before missing values were introduced. We applied the LGMI algorithm to each generated data set, creating 10 imputed data sets at each of 100 simulations run. An imputation at each simulation was considered completed when the convergence criteria in equation (7.1.3) was satisfied or after 100 attempts.

$$\left|\rho_{12} - \rho_{12}^{imp}\right| < c_{12}\rho_{12} \tag{7.1.3}$$

for ρ_{12} being the true correlation estimate, ρ_{12}^{imp}, the correlation estimate associated with the imputed data set, and c_{12} being some constant chosen to minimize standardized bias and maximize coverage rate calculated after all 100 simulations were run.

We also obtained the Pearson correlation for each imputed data set and then estimated the average correlation across the 10 imputed correlations at each simulation. We then calculated the average estimate (AE), standardized bias (SB), root mean square error (RMSE), coverage rate (CR), and average width of confidence intervals (AW) using these 100 estimates. Results involving pairwise correlations from application of the Lurie-Goldberg multiple imputation algorithm to the N(0,1), t_3, and Gamma(1,1) distributed data are given in Tables VI, VII, and VIII, respectively. True means for all generated data sets and AE values for Y_2, μ_2, for the data set are given in Table IX. Results from the naïve approach of imputing data directly via joint modeling under the normality assumption and from complete-case (CC) analyses are also shown. Here, estimates for pairwise correlations from complete-case analyses were less comparable to true pairwise correlations than those obtained from either imputation method, but were still reasonable.

With the new LGMI approach for all three distributions, the average estimate of the Pearson correlation for each generated data set is comparable to that from the original data set. The SB values are acceptable as they are all < 50%, the small RMSE values indicate quite good precision and accuracy, and the AW values furthermore are comparable to the 95% confidence interval widths of the true estimates. The coverage rate approximates 95%, additionally showing the validity of the LGMI algorithm in the bivariate continuous case. Likewise, we observe generally AE values of μ_2 comparable to true μ_2 obtained from the generated data (Table IX) among all data sets associated with any distribution or pairwise correlation. The naïve approach contrarily led to SB values > 50% and even SB values > 100% in several cases as well as CR estimates < 90% in several cases for *t* and Gamma distributed data as well as normally distributed data. Also, AW values appear to be artificially narrower than expected. Therefore, we observe that our approach could be a favorable alternative for imputing bivariate continuous data in MCAR cases.

Table VI: SIMULATION RESULTS FOR GENERATED N(0, 1) DATA

New LGMI Approach							
Data generated under the MCAR mechanism (50% missing)							Convergence Criteria Constant (Multiplied to ρ from generated data)
Data Set	True ρ from generated data	AE	SB	RMSE	CR	AW	
1	-0.8019	-0.8017	11.2524	0.0019	94.8498	0.14552	0.0125
2	-0.7977	-0.7975	9.7569	0.0018	95.5888	0.14813	0.0125
3	-0.3994	-0.3992	5.0815	0.0037	95.8267	0.335	0.0500
4	-0.3977	-0.3973	11.6103	0.0038	96.1692	0.3353	0.0500
5	0.4005	0.4014	22.2153	0.0039	95.8456	0.33434	0.0500
6	0.4010	0.4019	25.5326	0.0037	95.7298	0.33413	0.0500
7	0.7985	0.7979	35.9097	0.0018	95.3605	0.14795	0.0125
8	0.7990	0.7998	42.7713	0.0020	94.9202	0.14672	0.0125
Naïve Approach							
Data Set	True ρ from generated data	AE	SB	RMSE	CR	AW	CC Estimate
1	-0.8019	-0.8028	47.4940	0.0001	92.0432	0.0514	-0.8100
2	-0.7977	-0.7979	6.4270	0.0007	95.8449	0.0641	-08020
3	-0.3994	-0.3988	25.7081	0.0007	98.0512	0.1480	-0.4019
4	-0.3977	-0.3978	2.5724	0.0033	88.0019	0.1539	-0.3984
5	0.4005	0.3986	45.7707	0.0035	87.8668	0.1535	0.3976
6	0.4010	0.3994	61.2199	0.0025	93.4326	0.1492	0.4028
7	0.7985	0.7982	68.5948	0.0012	92.7341	0.0642	0.8040
8	0.7990	0.7989	8.5705	0.0011	92.8716	0.0640	0.8033

Table VII: SIMULATION RESULTS FOR GENERATED t_3 DATA

New LGMI Approach							
Data generated under the MCAR mechanism (50% missing)							Convergence Criteria Constant (Multiplied to ρ from generated data)
Data Set	True ρ from generated data	AE	SB	RMSE	CR	AW	
1	-0.8156	-0.8149	33.2935	0.0021	94.9074	0.13715	0.0125
2	-0.7768	-0.7764	22.6317	0.0016	96.1201	0.16116	0.0125
3	-0.4036	-0.4034	4.6243	0.0037	95.7791	0.33376	0.0500
4	-0.3963	-0.3952	29.7284	0.0038	95.8542	0.33602	0.0500
5	0.3870	0.3878	24.9229	0.0035	96.1759	0.3381	0.0500
6	0.4129	0.4113	37.8781	0.0044	95.2750	0.33124	0.0500
7	0.7781	0.7779	12.1033	0.0019	95.4169	0.16025	0.0125
8	0.8022	0.8018	22.6672	0.0018	94.8835	0.14549	0.0125
Naïve Approach							
Data Set	True ρ from generated data	AE	SB	RMSE	CR	AW	CC Estimate
1	-0.8156	-0.8183	161.6786	0.0024	89.4830	0.0586	-0.8188
2	-0.7768	-0.7754	102.0512	0.0019	91.3718	0.0707	-0.7738
3	-0.4036	-0.4041	1.7479	0.0026	90.9860	0.1508	-0.4076
4	-0.3963	-0.4014	110.2944	0.0062	84.8766	0.1529	-0.4004
5	0.3870	0.3874	24.2366	0.0015	95.5887	0.1504	0.3862
6	0.4129	0.4119	36.2520	0.0027	90.6629	0.1499	0.4157
7	0.7781	0.7767	73.4012	0.0018	90.2990	0.0704	0.7782
8	0.8022	0.8035	91.1763	0.0018	90.7876	0.0514	0.8042

Table VIII: SIMULATION RESULTS FOR GENERATED GAMMA(1,1) DATA

New LGMI Approach							
Data generated under the MCAR mechanism (50% missing)							Convergence Criteria Constant (Multiplied to ρ from generated data)
Data Set	True ρ from generated data	AE	SB	RMSE	CR	AW	
1	-0.8083	-0.8076	37.0654	0.0020	94.5748	0.1418	0.0125
2	-0.7917	-0.7917	1.8075	0.0018	95.4725	0.1519	0.0125
3	-0.4034	-0.4037	10.1429	0.0031	96.3221	0.3337	0.0500
4	-0.4004	-0.4003	3.9241	0.0042	95.7923	0.3346	0.0500
5	0.3823	0.3812	33.2641	0.0035	96.5697	0.3400	0.0500
6	0.4048	0.4046	7.3126	0.0035	96.1474	0.3335	0.0500
7	0.7838	0.7833	24.5795	0.0017	96.0341	0.1569	0.0125
8	0.8132	0.8124	38.8164	0.0022	94.0330	0.1387	0.0125
Naïve Approach							
Data Set	True ρ from generated data	AE	SB	RMSE	CR	AW	CC Estimate
1	-0.8083	-0.8104	156.6513	0.0025	89.7548	0.0609	-0.8061
2	-0.7917	-0.7910	69.8911	0.0014	92.5647	0.0664	-0.7894
3	-0.4034	-0.4021	22.2986	0.0034	89.3146	0.1511	-0.4074
4	-0.4004	-0.4005	13.5156	0.0028	90.5144	0.1514	-0.3959
5	0.3823	0.3813	19.2015	0.0031	89.7804	0.1540	0.3842
6	0.4048	0.4011	105.2134	0.0005	87.8566	0.1510	0.4023
7	0.7838	0.7832	51.3690	0.0014	92.3429	0.0686	0.7796
8	0.8132	0.8125	29.6884	0.0015	91.0094	0.0603	0.8127

Table IX: TRUE MEANS OF GENERATED DATA, COMPLETE-CASE (CC) ESTIMATES, AND AVERAGE ESTIMATES (AE) OF MEANS FOR THE IMPUTED VARABLE FOR Y_2, μ_2, FOR ALL GENERATED DATA SETS IN THE BIVARIATE CONTINUOUS CASE

LGMI Approach									
	N(0,1) data generated under MCAR mechanism			t_3 data generated under MCAR mechanism			Gamma(1,1) data generated under MCAR mechanism		
Data Set	True μ_1	True μ_2	AE of μ_2	True μ_1	True μ_2	AE of μ_2	True μ_1	True μ_2	AE of μ_2
1	-0.0066	-0.0682	-0.1372	-0.1218	0.1330	0.0961	1.0183	1.1094	1.0794
2	-0.0261	0.0094	-0.0117	-0.0319	-0.0570	-0.0421	1.0249	1.0632	1.0343
3	0.0492	-0.0980	-0.0811	0.0566	0.0173	-0.0341	1.0956	1.1090	1.0643
4	0.1772	-0.2332	-0.2554	-0.0389	-0.0482	0.0325	1.0458	1.0555	1.0610
5	-0.0971	-0.0099	-0.0024	0.0319	0.1721	0.1906	1.0570	1.0598	1.0624
6	-0.1400	-0.2502	-0.2801	-0.0223	0.0334	0.0328	0.9995	1.0912	1.1065
7	0.0308	0.1149	0.0750	-0.0610	-0.0321	-0.0608	1.0504	1.1353	1.0795
8	0.1606	0.1474	0.1741	0.0162	0.0027	-0.0239	1.0040	1.1287	1.1446
μ_2 Estimates based on Naïve and Complete Case Approaches									
Data Set	Naïve Approach		CC Approach	Naïve Approach		CC Approach	Naïve Approach		CC Approach
1	0.0362		-0.0979	-0.0307		0.1492	1.0015		0.9992
2	0.0426		-0.0351	0.1179		0.1102	1.0257		0.9723
3	0.1030		0.0458	-0.1161		0.0603	0.9831		1.1028
4	0.0560		-0.0502	-0.0728		-0.2276	0.9645		1.0520
5	0.0690		-0.0283	-0.0859		-0.0258	1.0037		1.1178
6	0.0354		-0.0111	-0.2030		0.0045	1.0076		1.0559
7	0.0268		-0.0684	-0.1063		0.1426	0.9537		1.1280
8	0.0628		0.0022	-0.3012		-0.0389	0.9280		1.2041

7.2 Multivariate Continuous Data

In these simulations, we extend our LGMI algorithm to MCAR multivariate data with $k = 3$ variables, Y_1, Y_2, and Y_3, where Y_1 is completely observed, 25% of Y_2 is missing and 25% of Y_3 is missing, such that 50% of data entries have observed values for all 3 variables and 50% of entries have a missing value in either Y_2 or Y_3, as shown in Figure 1, where shaded regions indicate entries with missing values.

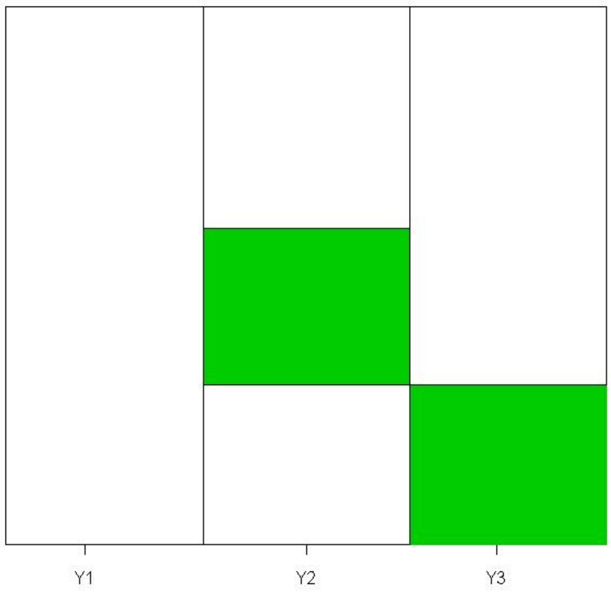

Figure 1: Plot showing three variables, Y_1, Y_2, and Y_3 with Y_2, and Y_3 having missing entries (shaded areas).

We first generated data under the N(0,1), t_3, or Gamma(1,1) distribution with correlations approximating -0.8, -0.4, 0.4, or 0.8. Each data set contained 200 variables and each variable of a particular data set followed the same distribution. True correlation and mean estimates were obtained before missingness was induced. We imposed missingness under the MCAR mechanism, where we randomly selected a subset comprising 25% of the data set in which we deleted Y_2 values and selected another subset of the data set in which we deleted Y_3 values. Therefore, 50% of entries in these data sets have a missing value in either Y_2 or Y_3.

We applied each data set to our LGMI algorithm involving 1000 iterations of $m = 10$ imputations. Each imputation was completed after the convergence criteria were satisfied or after 100 attempts at convergence were tried. The convergence of this algorithm was determined in the absolute difference between all $k(k-1)/2$, or 3, pairwise correlations of the applied data and of the imputed data were less than the pairwise correlations of the applied data multiplied by some constant c_{jk}, $j = 1,2$, $k = 2,3$, specified by the user, such that:

$$\begin{aligned} \left|\rho_{12} - \rho_{12}^{imp}\right| &< c_{12}\rho_{12}, \\ \left|\rho_{13} - \rho_{13}^{imp}\right| &< c_{13}\rho_{13}, \\ \left|\rho_{23} - \rho_{23}^{imp}\right| &< c_{23}\rho_{23} \end{aligned} \qquad (7.2.1)$$

In our simulations, convergence constants inputted into the algorithm were chosen for each data set to minimize standardized biases and maximize coverage rates. Table X gives the results for six data sets, including three distributions and two correlation matrices. Results give the true pairwise correlations, the pairwise correlations from the imputed data, the standardized bias (SB), root mean square error (RMSE), coverage rate (CR), and 95% confidence interval average width (AW) estimates for the correlations of the imputed data. As with bivariate continuous data, our LGMI approach was associated with AE values comparable to the true estimates, SB values < 50%, small RMSE values implying good precision and accuracy, CR values > 90%, and AW estimates comparable to confidence interval widths for the true pairwise correlations obtained from the generated data. Table XI gives the results from applying the joint modeling approach under the normality assumption directly to the multivariate continuous data and the complete-case results. Here, we observed that pairwise correlation estimates obtained from complete-case analyses were less comparable of true pairwise correlations than those obtained from either imputation method, albeit still reasonable.

Likewise, we observed SB values > 50% for pairwise correlations in some cases of normally, *t*, and Gamma distributed data, being more prominent given *t* and Gamma distributions. Lastly, overly optimistic coverage rates of 100% computed could be associated with AW estimates considerably larger than the confidence interval widths for the original estimates. Thus, we again infer that our method for imputing continuous data is a preferable alternative to the naïve approach of directly imputing data via joint modeling under the normality assumption.

Table X: RESULTS FROM APPLYING THE LGMI ALGORITHM TO MULTIVARIATE CONTINUOUS DATA GENERATED UNDER THE MCAR MECHANISM

Data Distribution	Coefficient	Convergence constant	True value	AE	SB	RMSE	CR	AW
N(0,1)	ρ_{12}	0.0075	0.8030	0.8028	11.27	0.00085	95.89	0.1008
N(0,1)	ρ_{13}	0.0075	-0.7984	-0.7986	20.06	0.00091	95.79	0.1026
N(0,1)	ρ_{23}	0.0075	-0.7831	-0.7825	42.01	0.00108	95.56	0.1098
			True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3
			0.0006	-0.0188	0.0370		-0.0507	0.0245
N(0,1)	ρ_{12}	0.0325	0.4080	0.4087	28.07	0.00207	95.49	0.2344
N(0,1)	ρ_{13}	0.0325	-0.4076	-0.4082	26.69	0.00216	95.55	0.2343
N(0,1)	ρ_{23}	0.0325	-0.4087	-0.4081	17.65	0.00254	94.10	0.2356
			True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3
			-0.0537	-0.0357	0.0628		-0.0063	0.0834
t_3	ρ_{12}	0.0075	0.7859	0.7862	25.81	0.00094	95.65	0.1084
t_3	ρ_{13}	0.0075	-0.7981	-0.7977	27.97	0.00125	94.28	0.1035
t_3	ρ_{23}	0.0075	-0.8068	-0.8073	36.70	0.00127	93.49	0.0994
			True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3
			-0.0369	-0.0120	-0.0412		0.0014	-0.1031
t_3	ρ_{12}	0.0275	0.4072	0.4066	22.70	0.00204	95.80	0.2342
t_3	ρ_{13}	0.0275	-0.3969	-0.3962	26.35	0.00205	95.66	0.2371
t_3	ρ_{23}	0.0275	-0.3871	-0.3874	9.97	0.00264	94.10	0.2397
			True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3
			0.2760	0.2607	-0.1549		0.2631	-0.1703
Gamma(1,1)	ρ_{12}	0.00925	0.7969	0.7962	39.21	0.00149	92.56	0.1048
Gamma(1,1)	ρ_{13}	0.00925	-0.7936	-0.7934	7.22	0.00138	93.05	0.1059
Gamma(1,1)	ρ_{23}	0.00925	-0.7845	-0.7844	6.06	0.00148	92.83	0.1100
			True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3
			0.8788	0.9863	1.1650		0.9770	1.1545
Gamma(1,1)	ρ_{12}	0.026975	0.3978	0.3962	47.74	0.00289	94.32	0.2375
Gamma(1,1)	ρ_{13}	0.026975	-0.4055	-0.4049	28.45	0.00180	96.25	0.2345
Gamma(1,1)	ρ_{23}	0.026975	-0.4021	-0.4010	36.15	0.00238	95.23	0.2360
			True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3
			0.9584	1.1705	0.9068		1.1190	0.8668

Table XI: RESULTS FROM APPLYING THE NAÏVE APPROACH OF IMPUTING DATA TO MULTIVARIATE CONTINUOUS DATA GENERATED UNDER THE MCAR MECHANISM

Data Distribution	Coefficient	True value	AE	SB	RMSE	CR	AW	Complete Case (CC) Analyses		
N(0,1)	ρ_{12}	0.8030	0.8018	77.46	0.00299	100.00	0.1439	0.8075	μ_2	0.1266
N(0,1)	ρ_{13}	-0.7984	-0.7943	185.17	0.00373	100.00	0.1486	-0.8073	μ_3	0.0476
N(0,1)	ρ_{23}	-0.7831	-0.7799	2.53	0.00180	100.00	0.1576	-0.8051		
		True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3			
		0.0006	-0.0188	0.0370		0.0432	0.0504			
N(0,1)	ρ_{12}	0.4080	0.4105	10.72	0.00375	100.00	0.3289	0.4125	μ_2	-0.0776
N(0,1)	ρ_{13}	-0.4076	-0.4061	45.93	0.00358	100.00	0.3304	-0.3920	μ_3	0.0089
N(0,1)	ρ_{23}	-0.4087	-0.4081	17.65	0.00254	100.00	0.3304	-0.4187		
		True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3			
		-0.0537	-0.0357	0.0628		-0.0765	-0.0813			
t_3	ρ_{12}	0.7859	0.7930	141.92	0.0003	100.00	0.1494	0.7933	μ_2	-0.0845
t_3	ρ_{13}	-0.7981	-0.8022	94.08	0.0027	100.00	0.1437	-0.7760	μ_3	0.1462
t_3	ρ_{23}	-0.8068	-0.8105	22.75	0.0018	100.00	0.1383	-0.8282		
		True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3			
		-0.0369	-0.0120	-0.0412		0.0093	0.0966			
t_3	ρ_{12}	0.4072	0.4022	156.65	0.0008	100.00	0.3316	0.4102	μ_2	0.2768
t_3	ρ_{13}	-0.3969	-0.4048	89.19	0.0057	100.00	0.3309	-0.4087	μ_3	-0.0964
t_3	ρ_{23}	-0.3871	-0.3917	26.03	0.0053	100.00	0.3305	-0.3998		
		True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3			
		0.2760	0.2607	-0.1549		0.3033	-0.0785			
Gamma(1,1)	ρ_{12}	0.7969	0.7951	90.57	0.0002	100.00	0.1481	0.8003	μ_2	0.9040
Gamma(1,1)	ρ_{13}	-0.7936	-0.7959	93.24	0.0022	100.00	0.1476	-0.8090	μ_3	0.9161
Gamma(1,1)	ρ_{23}	-0.7845	-0.7867	82.88	0.0022	100.00	0.1534	-0.7653		
		True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3			
		0.8788	0.9863	1.1650		0.9233	0.8835			
Gamma(1,1)	ρ_{12}	0.3978	0.3965	31.64	0.0004	100.00	0.3332	0.3947	μ_2	1.2595
Gamma(1,1)	ρ_{13}	-0.4055	-0.4094	79.52	0.0046	100.00	0.3293	-0.4119	μ_3	0.8644
Gamma(1,1)	ρ_{23}	-0.4021	-0.3998	53.83	0.0038	100.00	0.3323	-0.3785		
		True μ_1	True μ_2	True μ_3		AE of μ_2	AE of μ_3			
		0.9584	1.1705	0.9068		1.2464	0.8243			

7.3 Bivariate Binary Case

In examining our method for the imputing binary data described in Section 6.2 for the bivariate case with missing entries in the second variable, we created data sets with two binary variables and 500 observations and either randomly deleted 250 entries in the second variable to introduce missing values under the MCAR mechanism with 50% missingness.

We applied our approach to create 10 imputed data sets for each generated data set at each of 1000 simulations, assessing the performance by looking at the average estimate (AE), standardized bias (SB), root mean-square error (RMSE), coverage rate (CR), and average width (AW). Original (true) proportions were obtained by calculating the means of the variables before missingness was imposed. True phi coefficients for the generated data involved in the imputation approach were also computed before introducing missing values. Table XII includes results from our imputation approach, the naïve approach for imputing binary data, and complete-case analyses for these data. We then applied our method to bivariate data sets with both variables having missing entries per the approach described in 6.2 for such cases. Here, we imposed missing values in both variables under the MCAR mechanism by randomly deleting 25% entries in each variable. Again, 1000 simulations involving 10 imputations each were run for each generated data set and results are presented in Table XIII from the three approaches tried. Both tables indicate the validity of our method for different correlated binary data sets, as given by AE values comparable to the original estimates, SB values < 50%, small RMSE values indicating good precision and accuracy, CR values > 90%, and AW values comparable to confidence interval widths of the original pairwise phi coefficients.

Generally, pairwise correlation estimates obtained from complete-case analysis and either imputation method were comparable to the true values in all examples. Applying the joint modeling approach for imputing binary values based on a multinomial or loglinear model assumption to all the data sets generated nevertheless led to SB values grossly exceeding 50% or CR values < 90% in several cases, a possible result of multinomial or loglinear model assumption violations. These results thus suggest our method as a preferable alternative in imputing bivariate binary data.

Table XII: SIMULATION RESULTS FOR IMPUTING BIVARIATE BINARY DATA MISSING IN THE SECOND VARIABLE

Results for New Approach				Naïve Results			CC Results
True δ	-0.6810	True p_1	0.4860	True δ	-0.6810	Imputed p_2	Estimated δ
AE	-0.6807	True p_2	0.5560	AE	-0.6784	0.5685	-0.6800
SB	20.2230	Imputed p_2	0.5515	SB	38.5481		Estimated p_2
RMSE	0.0012	Convergence		RMSE	0.0036		0.5440
CR	94.7196	Constant	0.00875	CR	83.0621		
AW	0.0950			AW	0.0987		
True δ	-0.3925	True p_1	0.5260	True δ	-0.3925	Imputed p_2	Estimated δ
AE	-0.3919	True p_2	0.4760	AE	-0.4107	0.4641	-0.4200
SB	23.8818	Imputed p_2	0.4784	SB	297.3799		Estimated p_2
RMSE	0.0022	Convergence		RMSE	0.0182		0.4880
CR	93.1242	Constant		CR	68.6487		
AW	0.1506		0.01375	AW	0.1591		
True δ	0.3480	True p_1	0.5000	True δ	0.3480	Imputed p_2	Estimated δ
AE	0.3495	True p_2	0.4440	AE	0.3373	0.4382	0.3313
SB	47.5660	Imputed p_2	0.4464	SB	146.1941		Estimated p_2
RMSE	0.0028	Convergence		RMSE	0.0011		0.4400
CR	91.7202	Constant	0.0175	CR	73.0652		
AW	0.1576			AW	0.1728		
True δ	0.8087	True p_1	0.4680	True δ	0.8087	Imputed p_2	Estimated δ
AE	0.8086	True p_2	0.4920	AE	0.8132	0.5083	0.8200
SB	5.4819	Imputed p_2	0.4967	SB	71.6534		Estimated p_2
RMSE	0.0012	Convergence		RMSE	0.0061		0.4960
CR	92.0251	Constant	0.00875	CR	66.2257		
AW	0.0615			AW	0.0639		

Table XIII: SIMULATION RESULTS FOR IMPUTING BIVARIATE BINARY DATA MISSING IN BOTH VARIABLES

Results for New Approach				Naïve Results			CC Results
True δ	-0.7099	True p_1	0.5200	True δ	-0.7099	Imputed p_1	Estimated δ
AE	-0.7102	Imputed p_1	0.5346	AE	-0.7049	0.5279	-0.7059
SB	24.4278	True p_2	0.5040	SB	130.94	Imputed p_2	Estimated p_1
RMSE	0.0012	Imputed p_2	0.4947	RMSE	0.0053	0.4954	0.5251
CR	94.5486	Convergence		CR	80.9265		Estimated p_2
AW	0.0876	Constant	0.0075	AW	0.0917		0.5110
True δ	-0.4085	True p_1	0.4840	True δ	-0.4085	Imputed p_1	Estimated δ
AE	-0.4094	Imputed p_1	0.4942	AE	-0.4087	0.5355	-0.4100
SB	40.1080	True p_2	0.4960	SB	5.5803	Imputed p_2	Estimated p_1
RMSE	0.0019	Imputed p_2	0.49 3	RMSE	0.0032	0.5154	0.4800
CR	94.4952	Convergence		CR	88.0971		Estimated p_2
AW	0.1477	Constant	0.01275	AW	0.1523		0.4824
True δ	0.4083	True p_1	0.5200	True δ	0.4083	Imputed p_1	Estimated δ
AE	0.4078	Imputed p_1	0.5165	E	0.4085	0.5487	0.4085
S	15.4593	True p_2	0.5280	SB	3.0333	Imputed p_2	Estimated p_1
RMSE	0.0024	Imputed p_2	0.5277	RMSE	0.0035	0.5225	0.5440
CR	92.0796	Convergence		CR	87.0784		Estimated p_2
AW	0.1492	Constant	0.0175	AW	0.1522		0.5307
True δ	0.7392	True p_1	0.5260	True δ	0.7392	Imputed p_1	Estimated δ
AE	0.7390	Imputed p_1	0.5156	AE	0.7391	0.5170	0.7300
SB	14.3974	True p_2	0.5000	SB	3.6963	Imputed p_2	Estimated p_1
RMSE	0.0011	Imputed p_2	0.4906	RMSE	0.0017	0.4854	0.5147
CR	94.3719	Convergence		CR	91.0058		Estimated p_2
AW	.0803	Constant	0.0075	AW	0.0808		0.4996

7.4 Multivariate Binary Case

Testing our method in the multivariate case with k = 3 variables, we generated binary data sets with 100 entries and induced a 25% missingness pattern under the MCAR mechanism in each variable. Under this mechanism, entries were randomly deleted separately in each variable such that each entry could have missing values in one, two, or all three variables. Applying our method for imputing binary data, we once more ran 1000 simulations, each involving m = 10 imputations and presented the results from these simulations in Table XIV for each generated data set, given our approach and the naïve approach of imputing data directly via joint modeling as well as complete-case (CC) analysis results. As before, we calculated phi coefficients and true proportion estimates from before missingness was induced. Convergence for each simulation with our approach was achieved when the absolute difference between each of the original pairwise correlations and the pairwise correlations obtained from the imputed data was less than some constant c_{jk}, with j = 1,2 and k = 2,3 such that $\left|\delta_{jk}^{imp} - \delta_{jk}\right| < c_{jk}$for all pairwise correlations. As in the bivariate binary case, with our method, we observed AE values comparable to true estimates, SB values < 50%, small RMSE values associated with adequate precision and accuracy, CR estimates > 90%, and AW values comparable to confidence interval widths of true estimates for pairwise phi coefficients. These results therefore show validity of the new method when applied to multivariate binary data missing under the MCAR mechanism. Generally, pairwise correlation estimates obtained from complete-case analysis, the naïve imputation approach, and our semi-paramteric imputation method were comparable to the true values. The naïve approach, however, led to SB values > 50%, RMSE estimates > 50%, and CR values < 90%, possibly due to multinomial or loglinear model assumption violations. Given these results, we again see that our method may be a preferable avenue for imputing binary data.

Table XIV: SIMULATION RESULTS FOR IMPUTING MULTIVARIATE BINARY DATA

Results for New Approach							
Pairs	Convergence Constant	True δ	Imputed δ	SB	RMSE	CR	AW
(1,2)	0.025	0.7502	0.7495	20.5445	0.0029	93.2664	0.1769
(1,3)	0.025	0.4275	0.4278	9.0098	0.0027	96.3341	0.3249
(2,3)	0.05	0.2721	0.2692	36.5658	0.0067	90.6657	0.3755
		True p_1	True p_2	True p_3	Imputed p_1	Imputed p_2	Imputed p_3
		0.4900	0.4500	0.5200	0.4922	0.4636	0.5059
Pairs	Convergence Constant	True δ	Imputed δ	SB	RMSE	CR	AW
(1,2)	0.0025	-0.7679	-0.7677	4.3744	0.0034	91.4058	0.1664
(1,3)	0.025	0.4419	0.4414	11.6331	0.0035	95.1700	0.3211
(2,3)	0.0325	-0.3793	-0.3786	12.9950	0.0042	94.4604	0.3419
		True p_1	True p_2	True p_3	Imputed p_1	Imputed p_2	Imputed p_3
		0.4500	0.4900	0.5200	0.4599	0.4766	0.5288
	Naïve and Complete case (CC) results for MCAR case (Naïve case estimates compared to true estimates)						
Pairs		CC δ	Imputed δ	SB	RMSE	CR	AW
(1,2)		0.7087	0.7548	44.0284	0.0100	75.9132	0.1818
(1,3)		0.4739	0.4715	287.0018	0.0464	67.0812	0.3356
(2,3)		0.3281	0.2811	266.5011	0.0473	66.5150	0.4036
		CC p_1	CC p_2	CC p_3	Imputed p_1	Imputed p_2	Imputed p_3
		0.4861	0.5063	0.5405	0.4768	0.4586	0.5380
Pairs		CC δ	Imputed δ	SB	RMSE	CR	AW
(1,2)		-0.7069	-0.7208	332.5671	0.0471	62.2276	0.2078
(1,3)		0.4553	0.4315	57.6124	0.0167	73.7353	0.3577
(2,3)		-0.3754	-0.3963	63.1527	0.0163	75.6711	0.3534
		CC p_1	CC p_2	CC p_3	Imputed p_1	Imputed p_2	Imputed p_3
		0.4857	0.4805	0.5513	0.4622	0.5025	0.5290

7.5 Bivariate Mixed Case

To test our method for imputing mixed data, we first generated bivariate data with one variable as continuous following a N(5,1), t_3 or Gamma(1,1) distribution including 500 entries. Correlations were induced by sorting specific proportions of both continuous and binary variables. Imposing missingness under the MCAR mechanism was accomplished by randomly deleting 25% of entries in both variables such that 35% - 50% of entries in the data sets had missing values in the continuous variable, Y_1, binary variable, Y_2, or both variables.

1000 simulations involving m = 10 imputations each involving our imputation method for mixed data were run and performance of this method as well as for the naïve method was assessed via SB, RMSE, CR, and AW for the pairwise point-biserial correlations for each generated data set. Tables XV, XVI, and XVII give examples of favorable results with the new method associated with data sets having continuous variables following the normal, *t*, and Gamma distributions, respectively. Results involving the naïve approach of imputing data directly via joint modeling under the general local model and from complete-cases analyses are also shown. True means of continuous variables, true estimates of probabilities for binary variables, and true pairwise point-biserial correlations were obtained before missingness was imposed. Assessment measures indicate the new method as satisfactory in imputing bivariate mixed data for several cases of data missing under the MCAR mechanism for different distributions associated with the continuous variable, while the naïve approach was associated with results involving SB values > 50%, with several of these values exceeding 100%, and certain CR estimates < 90%, particularly in data with the continuous variable following a *t* or Gamma distribution. Furthermore, AE values obtained from our imputation approach were more comparable to the true pairwise correlations than were average estimates from the naïve approach or from complete-case analyses. These results may indicate violations of distributional and general location model assumptions, in which case our approach for imputing mixed data may be an attractive alternative.

Table XV: SIMULATION RESULTS FOR IMPUTING BIVARIATE MIXED DATA INVOLVING A N(5,1) DISTRIBUTION FROM IMPUTATION APPROACHES AND COMPLETE CASE (CC) ANALYSES

Results from New Approach				Results from Naïve Approach			
True δ	-0.7014	True μ_1	4.9226	True δ	-0.7014	Imputed μ_1	4.9874
AE	-0.7011	True p_2	0.4760	AE	-0.7072	Imputed p_2	0.5094
SB	14.3375	Imputed μ_1	4.9410	SB	195.2356		
RMSE	0.0014	Imputed p_2	0.4807	RMSE	0.0032		
CR	93.9152	Convergence	0.01	CR	90.3919		
AW	0.0901	Constant		AW	0.0885		
CC δ	-0.7177	CC μ_1	4.8974	CC p_2	0.5360		
True δ	-0.3799	True μ_1	5.0317	True δ	-0.3799	Imputed μ_1	5.0100
AE	-0.3800	True p_2	0.5520	AE	-0.3803	Imputed p_2	0.5193
SB	3.6626	Imputed μ_1	5.0259	SB	24.1380		
RMSE	0.0014	Imputed p_2	0.5494	RMSE	0.0016		
CR	95.7150	Convergence	0.01	CR	95.2425		
AW	0.1513	Constant		AW	0.1512		
CC δ	-0.4029	CC μ_1	5.0097	CC p_2	0.5400		
True δ	0.4232	True μ_1	5.0481	True δ	0.4232	Imputed μ_1	5.0043
AE	0.4239	True p_2	0.5320	AE	0.4241	Imputed p_2	0.4825
SB	41.0072	Imputed μ_1	5.0498	SB	48.3352		
RMSE	0.0015	Imputed p_2	0.4987	RMSE	0.0017		
CR	95.5342	Convergence	0.01	CR	95.1778		
AW	0.1451	Constant		AW	0.1450		
CC δ	0.4293	CC μ_1	5.0376	CC p_2	0.4840		
True δ	0.7164	True μ_1	5.1496	True δ	0.7164	Imputed μ_1	5.0320
AE	0.7161	True p_2	0.4800	AE	0.7183	Imputed p_2	0.4860
SB	14.8059	Imputed μ_1	5.1531	SB	110.1816		
RMSE	0.0014	Imputed p_2	0.4917	RMSE	0.0021		
CR	93.0934	Convergence	0.01	CR	92.9260		
AW	0.0864	Constant		AW	0.0857		
CC δ	-0.7177	CC μ_1	5.0804	CC p_2	0.4840		

Table XVI: SIMULATION RESULTS FOR IMPUTING BIVARIATE MIXED DATA INVOLVING A t_3 DISTRIBUTION FROM IMPUTATION APPROACHES AND COMPLETE CASE (CC) ANALYSES

Results from New Approach				Results from Naïve Approach			
True δ	-0.6465	True μ_1	-0.0827	True δ	-0.6465	Imputed μ_1	0.0582
AE	-0.6470	True p_2	0.4880	AE	-0.6263	Imputed p_2	0.4678
SB	30.1371	Imputed μ_1	-0.0621	SB	465.3172		
RMSE	0.0016	Imputed p_2	0.5060	RMSE	0.0202		
CR	93.5712	Convergence	0.01	CR	69.9672		
AW	0.1031	Constant		AW	0.1116		
CC δ	-0.6527	CC μ_1	0.0766	CC p_2	0.4630		
True δ	-0.3761	True μ_1	-0.1206	True δ	-0.3761	Imputed μ_1	-0.0601
AE	-0.3756	True p_2	0.4560	AE	-0.3645	Imputed p_2	0.4750
SB	26.0926	Imputed μ_1	-0.1127	SB	212.1641		
RMSE	0.0015	Imputed p_2	0.4553	RMSE	0.0117		
CR	95.7219	Convergence	0.01	CR	76.1424		
AW	0.1519	Constant		AW	0.1647		
CC δ	-0.3864	CC μ_1	-0.0399	CC p_2	0.4770		
True δ	0.3272	True μ_1	-0.1285	True δ	0.3272	Imputed μ_1	0.1044
AE	0.3273	True p_2	0.4920	AE	0.3170	Imputed p_2	0.4539
SB	6.6861	Imputed μ_1	-0.1251	SB	152.5677		
RMSE	0.0014	Imputed p_2	0.4898	RMSE	0.0107		
CR	96.1203	Convergence	0.01	CR	75.0604		
AW	0.1578	Constant		AW	0.1720		
CC δ	0.3259	CC μ_1	0.0878	CC p_2	0.4530		
True δ	0.6500	True μ_1	-0.0446	True δ	0.6500	Imputed μ_1	0.0002
AE	0.6504	True p_2	0.4760	AE	0.6423	Imputed p_2	0.4933
SB	24.0721	Imputed μ_1	-0.0966	SB	387.1463		
RMSE	0.0014	Imputed p_2	0.4539	RMSE	0.0008		
CR	94.1362	Convergence	0.01	CR	86.2967		
AW	0.1022	Constant		AW	0.1043		
CC δ	0.6558	CC μ_1	-0.0152	CC p_2	0.4870		

Table XVII: SIMULATION RESULTS FOR IMPUTING BIVARIATE MIXED DATA INVOLVING A GAMMA(1,1) DISTRIBUTION FROM IMPUTATION APPROACHES AND COMPLETE CASE (CC) ANALYSES

Results from New Approach				Results from Naïve Approach			
True δ	-0.6493	True μ_1	0.9801	True δ	-0.6493	Imputed μ_1	0.9637
AE	-0.6502	True p_2	0.5400	AE	-0.6349	Imputed p_2	0.5053
SB	46.5629	Imputed μ_1	0.9848	SB	579.0685		
RMSE	0.0016	Imputed p_2	0.5346	RMSE	0.01442		
CR	93.9631	Convergence	0.01	CR	74.1656		
AW	0.1023	Constant		AW	0.1066		
CC δ	-0.6550	CC μ_1	0.9633	CC p_2	0.491		
True δ	-0.3946	True μ_1	0.9122	True δ	-0.3946	Imputed μ_1	1.0561
AE	-0.3945	True p_2	0.4280	AE	-0.3828	Imputed p_2	0.4690
SB	8.6858	Imputed μ_1	0.9021	SB	404.6109		
RMSE	0.0014	Imputed p_2	0.4455	RMSE	0.01182		
CR	95.8202	Convergence	0.01	CR	84.4113		
AW	0.1493	Constant		AW	0.1535		
CC δ	-0.3901	CC μ_1	1.0567	CC p_2	0.4680		
True δ	0.4023	True μ_1	0.9315	True δ	0.4023	Imputed μ_1	0.9221
AE	0.4026	True p_2	0.5000	AE	0.3909	Imputed p_2	0.4830
SB	17.1405	Imputed μ_1	0.9286	SB	185.4703		
RMSE	0.0015	Imputed p_2	0.5222	RMSE	0.0115		
CR	95.5386	Convergence	0.01	CR	76.6835		
AW	0.1482	Constant		AW	0.1602		
CC δ	0.4033	CC μ_1	0.9330	CC p_2	0.4870		
True δ	0.6472	True μ_1	1.0071	True δ	0.6472	Imputed μ_1	1.0278
AE	0.6467	True p_2	0.5000	AE	0.6463	Imputed p_2	0.4685
SB	27.1875	Imputed μ_1	0.9672	SB	46.2167		
RMSE	0.0015	Imputed p_2	0.4890	RMSE	0.0017		
CR	93.9777	Convergence	0.01	CR	93.1569		
AW	0.1031	Constant		AW	0.1036		
CC δ	0.6580	CC μ_1	0.9784	CC p_2	0.4850		

7.6 Multivariate Mixed Case

We next tested our approach for multivariate mixed data by generating three types of data sets, each with 100 entries. Namely, we considered a trivariate data with two continuous variables and one binary variable, a trivariate data with one continuous variable and two binary variables, and a four-variable data set with two continuous variables and two binary variables. All the continuous variables in each particular generated data set followed the same distribution, which was either a normal, Gamma or mixture Gamma, or *t* distribution. Correlations were induced via sorting specific proportions in each continuous and binary variable. For the trivariate data sets, we generated 25% missing entries via the MCAR mechanism in each variable. Values were randomly deleted separately in each variable under the MCAR mechanism, leading to entries with missing values in one, two, or all three variables. In the 4-variable case, with the first two variables, Y_1 and Y_2, as continuous and the last two variables, Y_3 and Y_4, as binary, we generated two situations, one with one continuous variable missing and one with one binary variable missing. In the first situation, Y_2 had 50% missing entries under the MCAR mechanism. The second situation involved Y_4 with 50% missing entries under the MCAR mechanism. Each MCAR case involved randomly deleting 50% of values in the variable to be imputed.

Each data set generated included 100 entries and our method for imputing mixed data with 10 imputations was applied to each data set at each of 1000 simulations. All three pairwise correlations were evaluated for the trivariate cases and the three pairwise correlations involving the variable with missing data were evaluated in the 4-variable case. True estimates of means of continuous and binary variables and all true pairwise correlations in each case were calculated before missingness was imposed. Convergence was assessed at each imputation within each simulation when the absolute difference between each of the three pairwise correlations obtained from the imputed data and the corresponding true pairwise correlations of the original data were less than some constants c_{jk}, c_{jl}, and c_{kl}, based on the j^{th}, k^{th}, and l^{th} variables involved.

AE, SB, RMSE, CR, and AW values for pairwise correlations were evaluated. Results are given in Tables XVIII and XIX for the trivariate cases and Tables XX and XXI for the 4-variable case from our method for imputing mixed data, the naïve approach of imputing mixed data, and complete-case (CC) analysis results. Favorable results associated with our method were again indicated by AE values comparable to true estimates, SB values < 50%, small RMSE values indicating satisfactory precision and accuracy, CR values > 90%, and AW estimates comparable to the confidence interval widths for true estimates for pairwise correlations. In contrast, the naïve approach involving the general location model led to SB values > 50% and RMSE values > 0.005, potentially indicating poor accuracy and precision. Possibly poor accuracy and precision could potentially in turn indicate questionability in other assessment measures, such in the overly optimistic CR values of 100% observed. Possible violations of general location model assumptions in these cases thus again makes our approach an attractive alternative for imputing mixed data. AE values obtained from the new imputation method were most comparable to the true parameters.

Table XVIII: IMPUTATION RESULTS FOR TRIVARIATE DATA WITH ALL VARIABLES HAVING MISSING ENTRIES (2 CONTINUOUS VARIABLES, 1 BINARY VARIABLE) GIVEN NEW AND NAÏVE IMPUTATION AND COMPLETE-CASE (CC) APPROACHES

Order of Correlations: (Y_1, Y_2), (Y_1, Y_3), (Y_2, Y_3); Order of Means: Y_1, Y_2, Y_3								
NEW METHOD								
TRUE Correlation	Imputed Correlation	SB	RMSE	CR	AW	TRUE	Imputed	Convergence
						Means	Means	Constant
N(5,1) results under MCAR mechanism								
-0.7690	-0.7696	14.0590	0.0035	91.0875	0.1653	5.0660	5.0441	0.0250
0.3473	0.3469	14.4543	0.0021	97.3934	0.3481	5.0540	4.9917	0.0125
-0.3330	-0.3312	38.2269	0.0038	95.2266	0.3546	0.4400	0.4519	0.0250
t_3 results under MCAR mechanism								
0.2446	0.2463	30.3369	0.0047	94.1530	0.3749	-0.0781	-0.1031	0.0325
-0.5479	-0.5453	47.5926	0.0049	92.7539	0.2827	0.2999	0.2199	0.0325
-0.4037	-0.4038	3.0272	0.0047	93.3736	0.3358	0.4900	0.4961	0.0325
.75*Gamma(5,1) + .25*Gamma(1,1) results under MCAR mechanism								
-0.3350	-0.3355	9.6256	0.0038	95.1524	0.3534	0.8804	0.8907	0.0250
-0.5228	-0.5216	25.4221	0.0039	94.1745	0.2917	3.7687	3.8261	0.0250
0.4941	0.4929	29.6774	0.0031	95.5946	0.3021	0.5400	0.5495	0.01975
NAÏVE METHOD								
	Imputed Correlation	SB	RMSE	CR	AW	Imputed Means	CC Correlation	CC Means
N(5,1) results under MCAR mechanism								
	-0.7553	213.8648	0.0148	100.00	0.1728	5.1327	-0.7380	5.1000
	0.3360	178.5666	0.0143	100.00	0.3546	4.8311	0.3138	4.8050
	-0.3206	213.8647	0.0105	100.00	0.3546	0.4973	-0.3262	0.4930
t_3 results under MCAR mechanism								
	0.2274	342.0662	0.0226	100.00	0.3738	-0.0458	0.1792	0.0920
	-0.5440	85.3543	0.0073	100.00	0.2801	0.2167	-0.5652	0.2490
	-0.4036	47.9696	0.0066	100.00	0.3315	0.5219	-0.3920	0.5310
.75*Gamma(5,1) + .25*Gamma(1,1) results under MCAR mechanism								
	-0.3421	81.6665	0.0085	100.00	0.3493	3.9022	-0.3749	3.9200
	-0.5142	68.0128	0.0084	100.00	0.2927	3.9402	-0.5038	4.0230
	0.4863	40.9174	0.0008	100.00	0.3034	0.4786	0.4646	0.4860

Table XIX: IMPUTATION RESULTS FOR TRIVARIATE DATA WITH ALL VARIABLES HAVING MISSING ENTRIES (1 CONTINUOUS VARIABLE, 2 BINARY VARIABLES) GIVEN NEW AND NAÏVE IMPUTATION AND COMPLETE-CASE (CC) APPROACHES

Order of Correlations: (Y_1, Y_2), (Y_1, Y_3), (Y_2, Y_3); Order of Means: Y_1, Y_2, Y_3								
NEW METHOD								
TRUE Correlation	Imputed Correlation	SB	RMSE	CR	AW	TRUE Means	Imputed Means	Convergence Constant
N(5,1) results under MCAR mechanism								
-0.3968	-0.3948	36.2180	0.0044	94.0132	0.3377	5.0230	5.0494	0.0250
-0.3999	-0.4001	4.9991	0.0040	94.3233	0.3357	0.5300	0.5287	0.0250
0.2859	0.2869	18.7734	0.0042	94.3767	0.3663	0.4400	0.4340	0.0250
t_3 results under MCAR mechanism								
-0.2179	-0.2155	48.1378	0.0043	94.9570	0.3795	-0.0724	-0.0257	0.02500
-0.3667	-0.3657	18.6737	0.0043	94.2560	0.3460	0.4800	0.4682	0.02675
0.3126	0.3121	9.6684	0.0043	94.3174	0.3603	0.5400	0.5590	0.02675
Gamma(5,1) results under MCAR mechanism								
-0.2991	-0.2975	27.0884	0.0047	93.5797	0.3648	5.4220	5.4124	0.0275
-0.3703	-0.3676	48.3907	0.0048	93.9201	0.3459	0.5200	0.5288	0.0275
0.2358	0.2382	42.6390	0.0049	93.8312	0.3769	0.5100	0.5098	0.0275
NAÏVE METHOD								
	Imputed Correlation	SB	RMSE	CR	AW	Imputed Means	CC Correlation	CC Means
N(5,1) results under MCAR mechanism								
	-0.3880	163.7644	0.0123	100.00	0.3363	5.0831	-0.3529	5.1500
	-0.3772	332.2770	0.0228	100.00	0.3392	0.5136	-0.3974	0.5160
	0.3024	171.0045	0.0125	100.00	0.3588	0.5089	0.3046	0.5090
t_3 results under MCAR mechanism								
	-0.2185	19.5893	0.0062	100.00	0.3755	-0.1061	-0.2040	-0.1222
	-0.3692	8.9038	0.0065	100.00	0.3418	0.4860	-0.3695	0.4930
	0.3193	102.1634	0.0109	100.00	0.3549	0.5212	0.3453	0.5210
Gamma(5,1) results under MCAR mechanism								
	-0.3086	115.0179	0.0096	100.00	0.3574	5.3651	-0.3191	5.4067
	-0.3560	194.6836	0.0142	100.00	0.3454	0.5185	-0.3184	0.5200
	0.2365	38.0662	0.0081	100.00	0.3724	0.5128	0.2194	0.5000

Table XX: IMPUTATION RESULTS FOR 4-VARIABLE WITH Y_2 HAVING MISSING DATA (2 CONTINUOUS VARIABLES, 2 BINARY VARIABLES) UNDER THE MCAR MECHANISM GIVEN NEW, NAÏVE, AND COMPLETE-CASE (CC) APPROACHES

NEW APPROACH							
N(5,1) results							
Pairs*	TRUE Value	Imputed Value	SB	RMSE	CR	AW	Convergence Constant
(Y_1,Y_2)	-0.4600	-0.4599	1.9899	0.0037	94.6823	0.3152	0.0275
(Y_2,Y_3)	-0.5287	-0.5275	27.0068	0.0038	94.2454	0.2892	0.0275
(Y_2,Y_4)	0.5718	0.5710	17.1174	0.0040	93.4073	0.2710	0.0275
	True Means (Y_1, Y_2, Y_3, Y_4)				Imputed Mean (Y_2)		
	4.7692	5.1290	0.5000	0.4100	5.1079		
t_3 results							
Pairs	TRUE Value	Imputed Value	SB	RMSE	CR	AW	Convergence Constant
(Y_1,Y_2)	-0.4661	-0.4680	33.0556	0.0047	93.2653	0.3136	0.0275
(Y_2,Y_3)	-0.3644	-0.3656	22.6069	0.0042	94.2554	0.3461	0.0275
(Y_2,Y_4)	0.1131	0.1129	3.4907	0.0044	94.5300	0.3929	0.0275
	True Means (Y_1, Y_2, Y_3, Y_4)				Imputed Mean (Y_2)		
	0.0623	0.1988	0.5200	0.4700	0.1919		
Gamma(5,1) results							
Pairs	TRUE Value	Imputed Value	SB	RMSE	CR	AW	Convergence Constant
(Y_1,Y_2)	-0.3237	-0.3242	10.6014	0.0041	94.6382	0.3570	0.0275
(Y_2,Y_3)	-0.5061	-0.5038	46.8724	0.0044	93.7595	0.2993	0.0275
(Y_2,Y_4)	0.4787	0.4797	21.9528	0.0041	94.1715	0.3082	0.0275
	True Means (Y_1, Y_2, Y_3, Y_4)				Imputed Mean (Y_2)		
	4.9002	4.7650	0.5100	0.4700	4.8727		

* Pairs of variables involved in correlations with imputed data

Table XX: IMPUTATION RESULTS FOR 4-VARIABLE WITH Y_2 HAVING MISSING DATA (2 CONTINUOUS VARIABLES, 2 BINARY VARIABLES) UNDER THE MCAR MECHANISM GIVEN NEW, NAÏVE, AND COMPLETE-CASE (CC) APPROACHES (continued)

NAÏVE APPROACH							
N(5,1) results							
Pairs*	CC Value	Imputed Value	SB	RMSE	CR	AW	Imputed Mean (Y_2)
(Y_1,Y_2)	-0.4571	-0.4459	193.5660	0.0144	100.00	0.3176	5.0302
(Y_2,Y_3)	-0.4939	-0.5356	72.0677	0.0078	100.00	0.2838	CC Mean (Y_2)
(Y_2,Y_4)	0.5317	0.5510	276.7278	0.0191	100.00	0.5510	5.0280
t_3 results							
Pairs*	CC Value	Imputed Value	SB	RMSE	CR	AW	
(Y_1,Y_2)	-0.4315	-0.4571	102.0824	0.0103	100.00	0.3139	-0.0481
(Y_2,Y_3)	-0.3750	-0.3603	4.5608	0.0067	100.00	0.3443	CC Mean (Y_2)
(Y_2,Y_4)	0.1262	0.1199	109.6051	0.0113	100.00	0.3885	0.0103
Gamma(5,1) results							
Pairs*	CC Value	Imputed Value	SB	RMSE	CR	AW	Imputed Mean (Y_2)
(Y_1,Y_2)	-0.3536	-0.3318	134.9465	0.0125	100.00	0.3517	4.8002
(Y_2,Y_3)	-0.4937	-0.5022	98.7755	0.0094	100.00	0.2972	CC Mean (Y_2)
(Y_2,Y_4)	0.4372	0.4740	67.9162	0.0083	100.00	0.3080	4.9744

* Pairs of variables involved in correlations with imputed data

Table XXI: IMPUTATION RESULTS FOR 4-VARIABLE WITH Y_4 HAVING MISSING DATA (2 CONTINUOUS VARIABLES, 2 BINARY VARIABLES) UNDER THE MCAR MECHANISM GIVEN NEW, NAÏVE, AND COMPLETE-CASE (CC) APPROACHES

NEW APPROACH							
N(5,1) results							
Pairs*	TRUE Value	Imputed Value	SB	RMSE	CR	AW	Convergence Constant
(Y_1,Y_4)	-0.2682	-0.2688	13.0201	0.0037	95.3095	0.3687	0.0250
(Y_2,Y_4)	0.4727	0.4741	31.8377	0.0037	94.7260	0.3098	0.0250
(Y_3,Y_4)	-0.2000	-0.1981	40.4397	0.0040	95.2119	0.3816	0.0250
	True Means (Y_1, Y_2, Y_3, Y_4)				Imputed Mean (Y_4)		
	5.0092	5.0667	0.5300	0.5700	0.5732		
t_2 results							
Pairs	TRUE Value	Imputed Value	SB	RMSE	CR	AW	Convergence Constant
(Y_1,Y_4)	-0.1838	-0.1848	19.5088	0.0043	94.6983	0.3840	0.0275
(Y_2,Y_4)	0.2535	0.2526	16.2653	0.0044	94.3841	0.3733	0.0275
(Y_3,Y_4)	-0.1143	-0.1138	10.8272	0.0039	94.9647	0.3923	0.0275
	True Means (Y_1, Y_2, Y_3, Y_4)				Imputed Mean (Y_4)		
	0.1479	-0.1563	0.5500	0.4600	0.4418		
Gamma(5,1) results							
Pairs	TRUE Value	Imputed Value	SB	RMSE	CR	AW	Convergence Constant
(Y_1,Y_4)	-0.2725	-0.2717	15.6002	0.0040	95.0574	0.3683	0.0250
(Y_2,Y_4)	0.4509	0.4520	22.7873	0.0037	94.6976	0.3179	0.0250
(Y_3,Y_4)	-0.1639	-0.1631	19.4543	0.0031	96.2999	0.3854	0.0250
	True Means (Y_1, Y_2, Y_3, Y_4)				Imputed Mean (Y_4)		
	4.9826	4.9647	0.4700	0.4800	0.5250		

* Pairs of variables involved in correlations with imputed data

Table XXI: IMPUTATION RESULTS FOR 4-VARIABLE WITH Y_4 HAVING MISSING DATA (2 CONTINUOUS VARIABLES, 2 BINARY VARIABLES) UNDER THE MCAR MECHANISM GIVEN NEW, NAÏVE, AND COMPLETE-CASE (CC) APPROACHES (continued)

NAÏVE APPROACH							
N(5,1) results							
Pairs*	CC Value	Imputed Value	SB	RMSE	CR	AW	Imputed Mean (Y_4)
(Y_1,Y_4)	-0.3028	-0.2723	141.4935	0.0131	100.00	0.3654	0.5786
(Y_2,Y_4)	0.4605	0.4721	25.5215	0.0072	100.00	0.3086	CC Mean (Y_4)
(Y_3,Y_4)	-0.1775	-0.1982	19.6458	0.0073	100.00	0.3790	0.5700
t_3 results							
Pairs*	CC Value	Imputed Value	SB	RMSE	CR	AW	Imputed Mean (Y_4)
(Y_1,Y_4)	-0.2019	-0.1798	2.5228	0.0067	100.00	0.3813	0.4616
(Y_2,Y_4)	0.2294	0.2475	31.4166	0.0064	100.00	0.3704	CC Mean (Y_4)
(Y_3,Y_4)	-0.1256	-0.1207	157.7780	0.0114	100.00	0.3880	0.4800
Gamma(5,1) results							
Pairs*	CC Value	Imputed Value	SB	RMSE	CR	AW	Imputed Mean (Y_4)
(Y_1,Y_4)	-0.2987	-0.2846	70.8607	0.0065	100.00	0.3625	0.4482
(Y_2,Y_4)	0.4747	0.4540	53.7078	0.0068	100.00	0.3149	CC Mean (Y_4)
(Y_3,Y_4)	-0.1697	-0.1660	87.8184	0.0073	100.00	0.3830	0.4400

* Pairs of variables involved in correlations with imputed data

8. SIMULATIONS DEVISED AROUND REAL DATA

8.1 Description of Real Data

To exemplify applications of our imputation methods to real data sets, we used subsets from the Prostate SPORE database and the NYC HANES database. The Prostate SPORE database includes 3,452 men as of 2011 and is part of the Specialized Program of Research Excellence in Prostate Cancer (Grant #: P50 Ca 090386), established in 2001. The goal of this program is to enable collaboration between basic scientists, clinicians, and statisticians that would lead to new approaches for prostate cancer prevention, diagnosis, and treatment. The database contains demographic, clinical, pathological, and other information from patients treated at Northwestern Memorial Hospital, NorthShore University Health System facilities, and the Jessie Brown VA Hospital.

The NYC HANES (New York City Health and Nutrition Survey) study, with a database of 1999 subjects including 1168 women and 831 men, is modeled after NHANES (the National Health and Nutrition Survey) and involves a population-based cross-sectional design with data first collected in 2004. Data involving demographic, clinical, and other information were collected via physical examination, laboratory tests, face-to-face interviews, and computer-assisted self-interviews. NYC HANES was established to examine the prevalence of certain diseases and the effect of demographic variables and environmental factors on the prevalence rates. This program was conducted by the New York City Department of Health and Mental Hygiene and supported by the National Center for Health Statistics.

The following subsections focus on the scientific reasoning behind the variables used in examples of applying our semi-parametric methods for imputing continuous, binary, and mixed data. Namely, we describe the relationship of variables in the Prostate SPORE database for examples of bivariate continuous, multivariate continuous, bivariate mixed, and multivariate mixed cases, and of variables in the NYC HANES database for examples of bivariate continuous, bivariate binary, and multivariate binary cases.

8.1.1 Backround for Prostate SPORE Variables

The variables and their percentage of missing information used in the Prostate SPORE database include: percentage of the prostate gland with cancer (15.0%) prostate weight obtained from transrectal ultrasound (39.2% - 64.8%, depending on the data set), prostate weight obtained from digital rectal examination (4.4%), percentage of biopsy cores staining positive for cancer (17.0% - 19.0%, depending on the data set), biopsy Gleason score (0.0%), number of biopsy cores staining positive for cancer (24.5%), cancer present in seminal nodes of prostate gland (0.5%), cancer present in margins of prostate gland (0.0%), and cancer present in peripheral nerves of prostate gland (5.5%). Invcstigators working in prostate cancer research have become interested in the association between these variables. For example, Loeb *et al.* (2005) and Iczkowski *et al*. (2011) discuss studies relating the percentage of cancer in the prostate gland and prostate size and weight obtained either by digital rectal examination or by transrectal ultrasound. The correlations between biopsy-related variables, such as the percentage of biopsy cores staining positive for cancer and biopsy Gleason score, and variables obtained from radical prostatectomy, as percentage of the prostate gland positive for cancer, and presence of cancer in margins, seminal vesicles, and peripheral nerves obtained from the removed prostate, are also of great interest to investigators. This interest relates to the biopsy being a preferable alternative to radical prostatectomy (Mazzuchelli *et al.*, 2005; Montironi *et al.*, 2008; Bill-Axelson *et al.*, 2011).

8.1.2 Backround for NYC HANES Variables

Variables and their percentage of missing information in examples pertaining to the NYC HANES database involve: total cholesterol (0.0%) and triglyceride (26.8%) levels in women, indicators for entering the mainland US (45.9%), insurance offered at main job (40.6% - 42.0%, depending on the data set), private insurance (0.0%), and herpes I (12.0%) in women, and indicators for high blood pressure (3.5%) and entering the mainland US (47.1%) in men.

In the bivariate continuous case, we look at the correlations between total cholesterol and triglyceride, an important aspect of cardiovascular disease, in 1031 women; cardiovascular disease is a popular research field in women's health, since it is now considered the primary cause of mortality in women (Stampfer *et al.*, 2000; McSweeney *et al*., 2003; Hsia *et al.*, 2010). Variables in the bivariate binary and multivariate binary cases involving the NYC HANES database are important in determining if insurance offered at the job is affected by the length of time after immigration in the US and if this insurance is the main source of private insurance in women. Additionally, it could be of interest to see if women with particular afflictions, as infectious diseases, have access to insurance. Another potential question for investigators could involve the association between the length of time after immigration in the US and health conditions, an example of which is given by our correlation between entering the mainland US after 1990 and high blood pressure in men from the NYC HANES database.

8.2 Bivariate Continuous Case

Descriptions of these data used in the bivariate continuous cases are summarized in the Table XXII and descriptive statistics are given in Table XXIII.

Table XXII: DESCRIPTIONS OF BIVARIATE CONTINUOUS REAL DATA SETS USED

Real Data Set	Database Source	N	Variable 1	Variable 2
1	NYC HANES (women)	1031	Total Cholesterol (TC)	Triglycerides (TG)
2	PROSTATE SPORE	755	Percent Cancer based on radical prostatectomy (PERCENTCA)	Trans Ultrasound Prostate Weight (TRUS)
3	PROSTATE SPORE	732	Digital Rectal Prostate Weight (DRE_PROSTATE_WT)	Trans Ultrasound Prostate Weight (TRUS_PROSTATE_WT)

Table XXIII: SUMMARY STATISTICS OF BIVARIATE CONTINUOUS REAL DATA SETS USED

	NYC HANES data (N = 1031)							
Variable	Mean	SD	Median	Min.	Max.	Q25	Q75	NA's
TC	194.70	38.27	191.00	87.00	351.00	167.00	218.00	
TR	109.70	62.63	91.00	31.00	445.00	67.00	134.00	276
	PSPORE data (N = 755)							
Variable	Mean	SD	Median	Min.	Max.	Q25	Q75	NA's
PERCENTCA	8.34	5.60	6.99	0.08	23.01	4.99	10.02	
TRUS_PROSTATE_WT	40.96	22.22	34.80	9.99	180.02	27.00	47.75	489
	PSPORE data (N = 732)							
Variable	Mean	SD	Median	Min.	Max.	Q25	Q75	NA's
DRE_PROSTATE_WT	34.96	13.74	35.00	9.00	103.00	30.00	40.00	33
TRUS_PROSTATE_WT	42.42	18.76	36.00	17.10	180.00	28.00	49.00	296

We applied the Lurie-Goldberg multiple imputation algorithm to each of these real data sets and to data generated under the MCAR mechanism which reflected the characteristics of the real data.   With these real data, we first applied the LGMI algorithm using *m* = 10 imputations at each of 1000 simulations and compared the average of the Pearson correlations obtained from the imputed data to the correlation for the original data via AE, SB, RMSE, CR, and AW values. The convergence criterion was set as the difference between the average and original correlations being within 2.5% of the original correlation for each real data set applied. These assessment measures show the validity of the LGMI algorithm in imputing real bivariate continuous data given AE values comparable to original estimates, SB values < 50%, small RMSE values from which we could infer adequate precision and accuracy, CR estimates > 90%, and AW estimates comparable to 95% confidence interval widths of original estimates for all pairwise Pearson correlations (Table XXIV). Furthermore, AE values of mean estimates from imputed data were comparable to original means given in Table XXIII.

We also generated 100 data sets having the same characteristics as the original data and applied the Lurie-Goldberg multiple imputation algorithm to set of these data sets using $m = 10$ imputations. We then calculated the mean, standard deviation, median, minimum, maximum, first quartile, and third quartile for each generated and imputed data set. Missingness in these data sets was induced under the MCAR mechanism by randomly deleting a percentage of entries separately for each data set equal to the percentages of missing values in the respective variables of the original data.

Table XXIV: CHARACTERISTICS AND IMPUTATION RESULTS INCLUDING ASSESSMENT MEASURES FOR PAIRWISE CORRELATIONS INVOLVING NYC HANES AND PROSTATE SPORE BIVARIATE CONTINUOUS DATA

<table>
<tr><td colspan="4">Data Set 1</td></tr>
<tr><td colspan="4">NYC HANES (Women)
N = 1031</td></tr>
<tr><td colspan="4">Variable 1: TC; No. Missing: 0 (0.0%)</td></tr>
<tr><td colspan="4">Variable 2: TR; No. Missing: 276 (26.8%)</td></tr>
<tr><td>Original ρ</td><td>0.3474</td><td>Original μ_2</td><td>109.70</td></tr>
<tr><td>AE</td><td>0.3475</td><td>Imputed μ_2</td><td>109.32</td></tr>
<tr><td>SB</td><td>3.9296</td><td></td><td></td></tr>
<tr><td>RMSE</td><td>0.0012</td><td></td><td></td></tr>
<tr><td>CR</td><td>95.7997</td><td>Convergence</td><td></td></tr>
<tr><td>AW</td><td>0.1101</td><td>Constant</td><td>0.0125</td></tr>
<tr><td>Data Set 2</td><td></td><td></td><td></td></tr>
<tr><td colspan="4">Data Set 2</td></tr>
<tr><td colspan="4">PROSTATE SPORE
N = 755</td></tr>
<tr><td colspan="4">Variable 1: PERCENTCA; No. Missing: 0 (0.0%)</td></tr>
<tr><td colspan="4">Variable 2: TRUS_PROSTATE_WT; No. Missing: 489 (64.8%)</td></tr>
<tr><td>Original ρ</td><td>-0.2808</td><td>Original μ_2</td><td>40.96</td></tr>
<tr><td>AE</td><td>-0.2811</td><td>Imputed μ_2</td><td>40.67</td></tr>
<tr><td>SB</td><td>23.9679</td><td></td><td></td></tr>
<tr><td>RMSE</td><td>0.0011</td><td></td><td></td></tr>
<tr><td>CR</td><td>96.6458</td><td>Convergence</td><td></td></tr>
<tr><td>AW</td><td>0.1148</td><td>Constant</td><td>0.0100</td></tr>
<tr><td>Data Set 3</td><td></td><td></td><td></td></tr>
<tr><td colspan="4">Data Set 3</td></tr>
<tr><td colspan="4">PROSTATE SPORE
N = 732</td></tr>
<tr><td colspan="4">Variable 1: TRUS_PROSTATE_WT; No. Missing: 33 (4.5%)</td></tr>
<tr><td colspan="4">Variable 2: DRE_PROSTATE_WT; No. Missing: 296 (40.4%)</td></tr>
<tr><td>Original ρ</td><td>0.5333</td><td>Original μ_1</td><td>34.96</td></tr>
<tr><td>AE</td><td>0.5329</td><td>Imputed μ_1</td><td>34.85</td></tr>
<tr><td>SB</td><td>34.7882</td><td>Original μ_2</td><td>42.42</td></tr>
<tr><td>RMSE</td><td>0.0013</td><td>Imputed μ_2</td><td>42.05</td></tr>
<tr><td>CR</td><td>94.4883</td><td>Convergence</td><td></td></tr>
<tr><td>AW</td><td>0.0900</td><td>Constant</td><td>0.0175</td></tr>
</table>

Promising results were found with these simulated data created to resemble the original real data on average under the MCAR mechanism. 1000 simulations were run and at each simulation, a new data set reflecting the characteristics of the original NYC HANES data set or the first Prostate SPORE data set. Tables XXV to XXVI give the ranges for the means, standard deviations, medians, minimums, maximums, first quartiles, and third quartiles for the generated data and average values of these quantities from 10 imputed data sets associated with each generated data set. The ranges of the statistics of the generated data overlapped with the statistics obtained from the original data (Table XXIII). The statistics for the imputed values are in turn comparable to those of the generated values, further indicating that the Lurie-Goldberg multiple imputation algorithm is a promising method for handling missing entries in certain types of continuous data.

Table XXV: SIMULATION RESULTS FOR GENERATED AND IMPUTED DATA BASED ON NYC HANES DATA SET 1

Generated Data from NYC HANES under the MCAR mechanism									
	Generated Mean.1	Generated SD.1	Generated Median.1	Generated Min.1	Generated Max.1	Generated Q25.1	Generated Q75.1	Generatd ρ	Imputed ρ
Min.	193.9	37.9	188.4	1.2	308.1	165.1	213.7	0.3374	0.3212
Q25	194.0	38.2	190.5	21.1	330.1	166.5	216.5	0.3412	0.3371
Median	194.1	38.3	191.0	61.0	342.7	166.9	217.3	0.3459	0.3442
Mean	194.1	38.3	191.0	58.2	337.3	166.9	217.6	0.3462	0.3452
Q75	194.1	38.5	191.5	91.7	347.2	167.3	218.5	0.3508	0.3530
Max.	194.2	38.9	193.8	113.9	350.9	168.9	220.5	0.3574	0.3740
	Generated Mean.2	Generated SD.2	Generated Median.2	Generated Min.2	Generated Max.2	Generated Q25.2	Generated Q75.2		
Min.	106.4	57.7	88.3	0.6	360.0	64.5	126.7		
Q25	108.2	60.8	90.5	11.8	423.1	66.0	131.4		
Median	109.0	62.1	90.9	19.4	442.4	66.4	133.3		
Mean	109.1	62.1	91.4	20.3	427.8	66.4	133.1		
Q75	109.9	63.4	92.5	31.3	444.0	67.0	135.0		
Max.	111.7	66.4	95.1	37.6	445.0	68.5	140.3		
	Imputed Mean.2	Imputed SD.2	Imputed Median.2	Imputed Min.2	Imputed Max.2	Imputed Q25.2	Imputed Q75.2		
Min.	106.4	57.6	87.9	0.5	360.0	64.6	126.9		
Q25	108.1	60.5	90.4	9.9	423.1	65.9	131.0		
Median	109.0	61.7	91.1	16.7	442.4	66.4	133.0		
Mean	109.0	61.9	91.4	17.3	427.8	66.4	133.0		
Q75	109.9	63.4	92.5	26.3	444.0	67.0	134.9		
Max.	111.8	65.9	96.3	34.5	445.0	68.3	138.9		

Table XXVI: SIMULATION RESULTS FOR GENERATED AND IMPUTED DATA BASED ON PROSTATE SPORE DATA SET 2

Generated Data from Prostate SPORE under the MCAR mechanism									
	Generated Mean.1	Generated SD.1	Generated Median.1	Generated Min.1	Generated Max.1	Generated Q25.1	Generated Q75.1	Generatd ρ	Imputed ρ
Min.	8.201	5.421	5.019	0.001	21.570	4.982	10.010	-0.2906	-0.3042
Q25	8.293	5.552	6.953	0.028	22.990	4.985	10.020	-0.2846	-0.2863
Median	8.326	5.597	6.992	0.059	23.000	4.985	10.020	-0.2801	-0.2789
Mean	8.327	5.589	6.760	0.056	22.960	4.986	10.060	-0.2803	-0.2793
Q75	8.358	5.623	7.002	0.086	23.010	4.986	10.020	-0.2756	-0.2720
Max.	8.478	5.697	8.015	0.116	23.020	4.992	11.990	-0.2706	-0.2585
	Generated Mean.2	Generated SD.2	Generated Median.2	Generated Min.2	Generated Max.2	Generated Q25.2	Generated Q75.2		
Min.	36.750	16.420	32.010	0.044	102.000	24.780	42.030		
Q25	39.970	19.830	34.000	2.919	126.100	26.940	46.890		
Median	40.420	21.130	34.490	6.180	143.600	27.000	47.010		
Mean	40.430	21.110	34.640	6.431	145.100	27.060	47.680		
Q75	40.950	22.450	35.380	10.001	165.500	27.030	48.590		
Max.	43.230	25.210	37.020	17.026	179.100	28.310	52.670		
	Imputed Mean.2	Imputed SD.2	Imputed Median.2	Imputed Min.2	Imputed Max.2	Imputed Q25.2	Imputed Q75.2		
Min.	36.630	16.140	32.310	0.019	102.000	24.780	41.630		
Q25	39.860	19.520	33.970	1.120	126.100	26.840	46.630		
Median	40.230	20.780	34.560	2.604	143.600	27.000	47.180		
Mean	40.260	20.770	34.620	2.940	145.100	27.030	47.610		
Q75	40.720	22.120	35.170	4.188	165.500	27.340	48.560		
Max.	43.010	24.630	36.920	10.025	179.100	28.300	52.780		

8.3 Multivariate Continuous Case

To examine the performance of the LGMI method on multivariate continuous real data, we considered a subset of 100 men from the Prostate SPORE database including variables for percent of biopsy cores staining positive for prostate cancer, percent of prostate cancer in the prostate gland removed via radical prostatectomy, and the Gleason score obtained from the prostate tissue biopsied. Information on the variables of these data are given in Table XXVII.

Table XXVII: VARIABLES IN 100 MEN IN THE PROSTATE SPORE DATABASE USED IN THE LGMI APPLICATION TO MULTIVARIATE CONTINUOUS DATA

Variable	Label	Number (Percent) Missing
1	Percent Biopsy Cores Positive for Prostate Cancer	17 (17.0%)
2	Percent Cancer in Prostate Gland	15 (15.0%)
3	Biopsy Gleason Score	0 (0.0%)

Applying the LGMI algorithm to this data set, we observe adequate performance of the LGMI algorithm, via AE values comparable to original pairwise correlation estimates, SB < 50%, small RMSE values indicating good precision and accuracy, CR > 90%, and AW estimates comparable to expected confidence interval widths for original estimates.

Table XXVIII: ORIGINAL ESTIMATES AND AVERAGE ESTIMATES AND ASSESSMENT MEASURES FROM THE LGMI APPLICATION TO MULTIVARIATE CONTINUOUS DATA FROM THE PROSTATE SPORE DATABASE

Pairs	Convergence Constant	Original correlation	Imputed correlation	SB	RMSE	CR	AW
(1,2)	0.0425	0.4895	0.4911	39.8642	0.0033	95.024	0.3041
(1,3)	0.0675	0.2384	0.2396	39.3788	0.0026	97.0299	0.3728
(2,3)	0.075	0.3101	0.3111	23.2196	0.0037	95.1808	0.3596
		Original μ_1	Original μ_2	Original μ_3	Imputed μ_1	Imputed μ_2	Imputed μ_3
		32.39	10.65	6.5	31.9	10.66	6.5

We also created 1000 data sets with characteristics mimicking the original data and imposed missing values via the MCAR mechanism, by randomly deleting a percentage of entries in the first two variables equal to the respective percentages given in Table XXVII. The LGMI algorithm was applied to each data set at each of 1000 simulations with m = 10 imputations. Results given as summary statistics overlap quite well between the generated and imputed data (Table XXIX), further implying adequate performance of the LGMI application to multivariate continuous data.

Table XXIX: SUMMARY STATISTICS FOR GENERATED AND IMPUTED DATA INVOLVING THE LGMI APPLICATION MULTIVARIATE CONTINUOUS DATA FROM THE PROSTATE SPORE DATABASE (MCAR CASE)

	Generated ρ_{12}	Generated ρ_{13}	Generated ρ_{23}	Imputed ρ_{12}	Imputed ρ_{13}	Imputed ρ_{23}
Min.	0.442	0.1791	0.266	0.4575	0.1976	0.2778
Q25	0.4755	0.2159	0.2952	0.4756	0.2146	0.2948
Median	0.4926	0.2321	0.3111	0.4921	0.2314	0.3097
Mean	0.4918	0.2319	0.3115	0.4912	0.2307	0.3105
Q75	0.5079	0.2484	0.3271	0.5073	0.2473	0.327
Max.	0.5656	0.2844	0.3671	0.5224	0.2624	0.3425
	Generated μ_1	Generated μ_2	Generated μ_3	Imputed μ_1	Imputed μ_2	Imputed μ_3
Min.	27.76	8.528	6.136	27.63	8.473	6.136
Q25	31.39	10.543	6.425	31.48	10.632	6.425
Median	32.38	11.128	6.49	32.43	11.151	6.49
Mean	32.39	11.120	6.495	32.47	11.17	6.495
Q75	33.38	11.661	6.566	33.49	11.706	6.566
Max.	36.33	13.934	6.812	36.41	14.239	6.812

8.4 Bivariate Binary Case

To examine the performance of our method for imputing binary data on a real data set, we considered three data sets from the NYC HANES database, two pertaining to women (N = 1168) and one pertaining to men (N = 831). The main characteristics of variables included in each data set are given in Table XXX. Results were obtained from 1000 simulation runs involving 10 imputations at each run. Assessment measures in Table XXX indicate that the method works fairly well when applied to these real data, with AE values from imputed data comparable to original estimates, SB values < 50%, small RMSE values indicating good precision and accuracy, CR estimates > 90%, and AW estimates comparable to 95% confidence intervals of original estimates for pairwise correlation coefficients.

1000 data sets were also generated reflecting the characteristics of the real data under the MCAR mechanism, where percentages of entries equal to the respective percentages given in the descriptions for the data sets in Table XXX were randomly deleted. Results in Table XXXI again indicate adequate performance of the method, via sufficient overlapping of summary statistics.

Table XXX: CHARACTERISTICS AND IMPUTATION RESULTS INCLUDING ASSESSMENT MEASURES FOR PAIRWISE CORRELATIONS INVOLVING NYC HANES BIVARIATE BINARY DATA

Data Set 1			
NYC HANES (Women) N = 1168			

Variable 1: Entered Mainland US (After vs. Before 1990);
No. Missing 536 (45.9%):			

Variable 2: Insurance offered at main job; No. Missing 474 (40.6%):			
Original δ	-0.2313	Original p_1	0.5111
AE	-0.2309	Imputed p_1	0.4888
SB	18.6631	Original p_2	0.5908
RMSE	0.0018	Imputed p_2	0.5596
CR	94.9140	Convergence	
AW	0.1679	Constant	0.0125
Data Set 2			
NYC HANES (Women) N = 1168			
Variable 1: Private Insurance; No. Missing: 0 (0.0%)			
Variable 2: Insurance at main job; No. Missing: 474 (40.6%)			
Original δ	0.3102	Original p_1	0.6447
AE	0.3110	Imputed p_1	.
SB	47.0573	Original p_2	0.5908
RMSE	0.0016	Imputed p_2	0.5442
CR	95.8405	Convergence	
AW	0.1596	Constant	0.0100
Data Set 3			
NYC HANES MEN; N = 831			
Variable 1: High Blood Pressure; No. Missing: 29 (3.5%):			

Variable 2: Entered Mainland US (After vs. Before 1990);
No. Missing: 391 (47.1%):			

Original δ	-0.2337	Original p_1	0.2369
AE	-0.2335	Imputed p_1	0.2375
SB	7.2563	Original p_2	0.5750
RMSE	0.0023	Imputed p_2	0.5556
CR	93.1430	Convergence	
AW	0.1689	Constant	0.0175

Table XXXI: SUMMARY STATISTICS FOR GENERATED AND IMPUTED DATA INVOLVING NYC HANES BIVARIATE BINARY DATA (MCAR CASE)

	Entered Mainland US vs. Insurance offered at Main Job (Women)					
	Generated δ	Generated p_1	Generated p_2	Imputed δ	Imputed p_1	Imputed p_2
Min.	-0.2500	0.4383	0.5389	-0.2977	0.4422	0.5247
Q25	-0.2400	0.4842	0.5807	-0.2413	0.4857	0.5665
Median	-0.2300	0.4984	0.5922	-0.2314	0.4987	0.5769
Mean	-0.2323	0.4986	0.5932	-0.2327	0.4989	0.5770
Q75	-0.2200	0.5111	0.6052	-0.2238	0.5115	0.5872
Max.	-0.2200	0.5712	0.6614	-0.2070	0.5564	0.6248
	Private Insurance vs. Insurance offered at Main Job (Women)					
	Generated δ	Generated p_1	Generated p_2	Imputed δ	Imputed p_1	Imputed p_2
Min.	0.3000	0.6156	0.5231	0.2852	0.6156	0.5252
Q25	0.3100	0.6498	0.5706	0.3055	0.6498	0.5696
Median	0.3200	0.6601	0.5821	0.3137	0.6601	0.5813
Mean	0.3135	0.6601	0.5825	0.3131	0.6601	0.5820
Q75	0.3200	0.6695	0.5951	0.3211	0.6695	0.5949
Max.	0.3200	0.7012	0.6398	0.3345	0.7012	0.6373
	High Blood Pressure vs. Entered Mainland US (Men)					
	Generated δ	Generated p_1	Generated p_2	Imputed δ	Imputed p_1	Imputed p_2
Min.	-0.2500	0.2020	0.5068	-0.3061	0.2059	0.5002
Q25	-0.2400	0.2394	0.5545	-0.2440	0.2444	0.5540
Median	-0.2300	0.2494	0.5705	-0.2344	0.2531	0.5693
Mean	-0.2343	0.2496	0.5692	-0.2348	0.2539	0.5685
Q75	-0.2200	0.2594	0.5818	-0.2248	0.2644	0.5824
Max.	-0.2200	0.3142	0.6295	-0.1693	0.3171	0.6300

8.5 Multivariate Binary Case

We next applied our method for imputing multivariate binary data to a subset of 200 women in the NYC HANES database including indicator variables for Herpes I, the offering of insurance at the workplace, and having private insurance. Information for these variables is given in Table XXXII.

Table XXXII: VARIABLES IN 200 WOMEN IN THE NYC HANES DATABASE USED IN THE APPLICATION FOR IMPUTING MULTIVARIATE BINARY DATA

Variable	Label	Number (Percent) Missing
1	Herpes I (yes vs. no)	12 (12.0%)
2	Insurance Offered at Workplace (yes vs. no)	42 (42.0%)
3	Private Insurance (yes vs. no)	0 (0.0%)

Creating 10 imputed data sets via our method for imputing mixed data at each of 1000 simulations and examining assessment measures given in Table XXXIII, we again see adequate performance of the new approach, via AE values comparable to original pairwise phi coefficient estimates, SB estimates < 50%, small RMSE values indicating adequate precision and accuracy, CR values > 90%, and AW estimates comparable to confidence interval widths for original estimates.

Table XXXIII: ORIGINAL ESTIMATES AND AVERAGE ESTIMATES AND ASSESSMENT MEASURES FROM THE IMPUTATION APPLICATION TO MULTIVARIATE BINARY DATA FROM THE NYC HANES (WOMEN) DATABASE

Pairs	Convergence Constant	Original δ	Imputed δ	SB	RMSE	CR	AW
(1,2)	0.0325	-0.1422	-0.1435	22.5165	0.0047	94.1559	0.3907
(1,3)	0.0325	-0.1376	-0.1388	23.3595	0.0040	95.0411	0.3900
(2,3)	0.0325	0.5131	0.5132	3.8314	0.0043	93.2211	0.2964
		Original p_1	Original p_2	Original p_3	Imputed p_1	Imputed p_2	Imputed p_3
		0.77	0.55	0.68	0.76	0.47	0.68

Similarly, satisfactory results were seen after generating 1000 data sets with characteristics mimicking those of the original data. For these generated data, missing values were induced under the MCAR mechanism, again by randomly deleting percentages of entries equal to the percentages of missing values in the original data. Table XXXIV gives the summary statistics associated with these generated and corresponding imputed data sets, again with 10 data sets imputed at each of 1000 simulations. Adequate performance of the imputation method is indicated by sufficient overlapping of the summary statistics.

Table XXXIV: SUMMARY STATISTICS FOR GENERATED AND IMPUTED DATA INVOLVING THE IMPUTATION APPLICATION TO MULTIVARIATE BINARY DATA FROM THE NYC HANES DATABASE (MCAR CASE – WOMEN)

	Generated δ_{12}	Generated δ_{13}	Generated δ_{23}	Imputed δ_{12}	Imputed δ_{13}	Imputed δ_{23}
Min.	-0.2007	-0.2022	0.4482	-0.2336	-0.2168	0.4069
Q25	-0.1606	-0.1895	0.4711	-0.1549	-0.1929	0.4801
Median	-0.1273	-0.1724	0.4964	-0.1220	-0.1753	0.5079
Mean	-0.1314	-0.1665	0.5017	-0.1229	-0.1703	0.5106
Q75	-0.1005	-0.1485	0.5297	-0.0933	-0.1511	0.5390
Max.	-0.0784	-0.0837	0.5720	-0.0183	-0.0791	0.6119
	Generated p_1	Generated p_2	Generated p_3	Imputed p_1	Imputed p_2	Imputed p_3
Min.	0.6426	0.5017	0.5300	0.6460	0.4800	0.5300
Q25	0.7751	0.5834	0.6538	0.7775	0.5725	0.6538
Median	0.7966	0.6100	0.6850	0.8013	0.5965	0.6850
Mean	0.7975	0.6098	0.6811	0.7992	0.5999	0.6811
Q75	0.8288	0.6384	0.7150	0.8285	0.6335	0.7150
Max.	0.9076	0.7067	0.7900	0.9120	0.7105	0.7900

8.6 Bivariate Mixed Case

We first applied the method for imputing mixed data to a bivariate subset of 200 men from the Prostate SPORE database, including a continuous variable of total number of biopsy cores positively staining for prostate cancer and binary variables indicating presence or absence of prostate cancer in the seminal vesicle removed, in the margins of the prostate gland removed, or in the nerves near the removed prostate. Table XXXV gives the characteristics of these data and also indicates the adequate performance of the method when applied to these data via assessment measures of AE values comparable to original estimates, SB estimates < 50%, small RMSE estimates associated with sufficient precision and accuracy, CR values > 90%, and AW estimates comparable to original pairwise correlation estimates. These assessment measures were obtained from 1000 simulations involving 10 imputed data sets at each simulation. Creating 1000 simulated data sets with these characteristics via MCAR mechanisms involved random deleting a percentage of values equal to the percentage of missing entries in each respective variable of the original data. The imputation method was then applied to each data set for m = 10 imputations. Here, we see that the range of summary statistics for the pairwise correlation estimates overlap quite well for data generated.

Table XXXV: CHARACTERISTICS AND IMPUTATION RESULTS INCLUDING ASSESSMENT MEASURES FOR PAIRWISE CORRELATIONS INVOLVING PROSTATE SPORE BIVARIATE MIXED DATA

Data Set 1			
PROSTATE SPORE N = 200			
Variable 1: Biopsy cores positive for Cancer; No. Missing: 49 (24.5%):			
Variable 2: Seminal vesicle (Positive vs. negative); No. Missing: 1 (0.5%):			
Original δ	0.0909	Original μ_1	2.9678
AE	0.0906	Imputed μ_1	2.9828
SB	20.0823	Original p_2	0.0452
RMSE	0.0015	Imputed p_2	0.0463
CR	96.0800	Convergence	
AW	0.1753	Constant	0.0100
Data Set 2			
PROSTATE SPORE N = 200			
Variable 1: Biopsy cores positive for Cancer; No. Missing: 49 (24.5%):			
Variable 2: Marginal nodes (Positive vs. negative); No. Missing: 0 (0.0%):			
Original δ	0.0960	Original μ_1	2.9678
AE	0.0968	Imputed μ_1	2.9936
SB	2.7382	Original p_2	0.1700
RMSE	0.0014	Imputed p_2	.
CR	96.2520	Convergence	
AW	0.1751	Constant	0.0100
Data Set 3			
PROSTATE SPORE N =200			
Variable 1: Biopsy cores positive for Cancer; No. Missing: 49 (24.5%):			

Variable 2: Peripheral nerves (Positive vs. negative);
No. Missing: 11 (5.5%):			

Original δ	0.270982	Original μ_1	2.9678
AE	0.270980	Imputed μ_1	2.9700
SB	0.0806	Original p_2	0.6614
RMSE	0.0015	Imputed p2	0.6593
CR	95.7429	Convergence	
AW	0.1711	Constant	0.0100

Table XXXVI: SUMMARY STATISTICS FOR GENERATED AND IMPUTED DATA INVOLVING PROSTATE SPORE BIVARIATE MIXED DATA (MCAR CASE)

	Positive Cores vs. Seminal Vesicle (+/-)					
	Generated δ	Generated μ_1	Generated p_2	Imputed δ	Imputed μ_1	Imputed p_2
Min.	0.0809	2.1970	0.0350	0.0623	1.6880	0.0320
Q25	0.0855	2.8430	0.0450	0.0854	2.3650	0.0420
Median	0.0902	3.0080	0.0500	0.0903	2.5330	0.0500
Mean	0.0905	3.0040	0.0535	0.0905	2.5210	0.0525
Q75	0.0956	3.1590	0.0600	0.0956	2.6770	0.0600
Max.	0.1009	3.7630	0.1050	0.1064	3.3200	0.1060
	Positive Cores vs. Marginal Nodes (+/-)					
	Generated δ	Generated μ_1	Generated p_2	Imputed δ	Imputed μ_1	Imputed p_2
Min.	0.0860	2.1820	0.0850	0.0811	1.6800	0.0840
Q25	0.0904	2.8330	0.1537	0.0908	2.3660	0.1515
Median	0.0955	3.0000	0.1700	0.0957	2.5210	0.1690
Mean	0.0955	2.9980	0.1699	0.0958	2.5220	0.1690
Q75	0.1002	3.1580	0.1850	0.1005	2.6810	0.1860
Max.	0.1060	3.8200	0.2450	0.1097	3.3450	0.2460
	Positive Cores vs. Peripheral Nerves (+/-)					
	Generated δ	Generated μ_1	Generated p_2	Imputed δ	Imputed μ_1	Imputed p_2
Min.	0.2610	2.3310	0.5556	0.2579	1.8460	0.5550
Q25	0.2653	2.8300	0.6349	0.2662	2.3580	0.6331
Median	0.2705	2.9970	0.6614	0.2711	2.5260	0.6584
Mean	0.2707	2.9990	0.6587	0.2714	2.5240	0.6568
Q75	0.2757	3.1550	0.6825	0.2765	2.6850	0.6804
Max.	0.2810	3.6750	0.7513	0.2912	3.3270	0.7516

8.7 Multivariate Mixed Case

In our final example, we investigated our imputation method for mixed data in the multivariate setting, by considering at a subset of 100 men from the Prostate SPORE database with two continuous variables, percent cancer in the prostate gland removed by radical prostatectomy and percent of biopsy needle cores staining positively for cancer, and one binary variable indicating the presence or absence of cancer in the marginal nodes. Here, 19% of the entries in the second variable of percent cancer in the prostate gland were missing, as shown in the variable descriptions given in Table XXXVII.

Table XXXVII: VARIABLES IN 100 MEN IN THE PROSTATE SPORE DATABASE USED IN THE IMPUTATION APPLICATION TO MULTIVARIATE MIXED DATA

Variable	Label	Number (Percent) Missing
1	Percent Cancer in Prostate Gland	0 (0.0%)
2	Percent Biopsy Cores Positive for Prostate Cancer	19 (19.0%)
3	Marginal nodes (positive vs. negative)	0 (0.0%)

Table XXXVIII gives on assessment measures of AE values comparable to original estimates, SB values < 50%, RMSE values sufficiently small associated with adequate precision and accuracy, CR estimates > 90%, and AW values comparable to confidence interval widths of original estimates for pairwise correlations involving the variable with missing data. From these results, we could therefore infer that our method performs adequately for this real multivariate mixed data set.

With generating 1000 data sets having characteristics of the original data and applying our method with 10 imputations to each data set at each of 1000 simulations, we also observe satisfactory results as indicated by the overlapping summary statistics. Table XXXIX gives the results with missingness in the generated data induced under the MCAR mechanism, created via randomly deleting 19% of entries in the second variable. With respect to overlapping summary statistics, we namely see comparable results between the means of generated and imputed data for Y_2 and between the pairwise correlations involving Y_2 associated with generated and imputed data.

Table XXXVIII: ORIGINAL ESTIMATES AND AVERAGE ESTIMATES AND ASSESSMENT MEASURES FROM THE IMPUTATION APPLICATION TO MULTIVARIATE MIXED DATA FROM THE PROSTATE SPORE DATABASE

Pairs	Convergence Constant	Original δ	Imputed δ	SB	RMSE	CR	AW
(1,2)	0.0275	0.3366	0.3362	8.6253	0.0035	95.3717	0.3530
(2,3)	0.0275	0.2640	0.2651	22.7556	0.0040	95.0477	0.3699
		Original μ_1	Original μ_2	Original p_1	Imputed μ_2		
		10.02	24.43	0.17	24.84		

Table XXXIX: SUMMARY STATISTICS FOR GENERATED AND IMPUTED DATA INVOLVING THE IMPUTATION APPLICATION TO MULTIVARIATE MIXED DATA FROM THE PROSTATE SPORE DATABASE (MCAR CASE)

	Generated δ_{12}	Generated δ_{13}	Generated δ_{23}	Imputed δ_{12}	Imputed δ_{23}
Min.	0.3041	0.2924	0.2316	0.2765	0.2137
Q25	0.3179	0.3081	0.2510	0.3303	0.2568
Median	0.3338	0.3239	0.2660	0.3360	0.2623
Mean	0.3344	0.3238	0.2657	0.3360	0.2623
Q75	0.3498	0.3380	0.2810	0.3419	0.2678
Max.	0.3690	0.3573	0.2965	0.3706	0.2829
	Generated μ_1	Generated μ_2	Generated p_3	Imputed μ_2	
Min.	8.5830	22.4100	0.3500	22.1400	
Q25	10.1880	26.3500	0.4700	26.4600	
Median	10.7610	27.8100	0.5000	27.8700	
Mean	10.7560	27.7200	0.5033	27.8100	
Q75	11.2880	29.1000	0.5400	29.2500	
Max.	13.3920	36.2400	0.6500	36.4300	

Through bivariate and multivariate real data examples, we have shown promising performance of our developed methods for imputing continuous, binary, and mixed data. We thus recommend the presented methods for missingness under the MCAR mechanism. Future work will involve extending our methods to data missing under any MAR mechanism.

8.8 Advantages of Multiple Imputation over Complete-Case Analyses in Real Data Applications

Although average estimates for our imputation methods are comparable to parameter estimates from complete-case analyses involving the real data presented in this chapter and in several cases of generated data presented in Chapter 7, multiple imputation can still be a preferable alternative to complete-case analyses for various reasons. As discussed in Section 2.3, for example, multiple imputation allows us to account for variation is the missing data, leading to potentially greater validity with respect to inferences (Schafer and Olsen, 1998). Multiple imputation also can be beneficial in that it can be used in fill in missing entries in unbalanced data. This application can allow for analyses and variance component estimation involving repeated measures ANOVA models, for instance, that may not converge otherwise when applied to unbalanced data (Laird and Ware, 1982; Hedeker and Gibbons, 2006). Lastly, a substantial fraction of missing information can lead to complete-case analyses of a data set of notably smaller sample size, potentially associated with a loss of power (Schafer, 1999). These situations therefore present examples involving real data applications, where we may choose multiple imputation rather than complete-case analyses even when parameter estimates obtained from the two approaches are comparable.

9. CONCLUSION

In this work, we have developed semi-parametric approaches for imputing continuous, binary, and mixed data. We have adopted principles of eCDF computation and the Lurie-Goldberg algorithm to impute continuous data (LGMI) and methods given in Emrich and Piedmonte (1991) and Demirtas and Doganay (2012) to impute binary data. We combined approaches for imputing continuous and binary data to impute mixed data. These discussed methods involve data transformations leading to values following the normal distribution that can be then imputed using joint modeling, constituting the parametric portion of our methods. Back-transformations via the Barton and Schruben (1993) method for continuous data and use of quantiles for binary variables then constitute the nonparametric portion of our methods. Simulations conducted on generated and real data indicate these techniques as promising in MCAR cases. Future work will include extending these methods to MAR scenarios. We therefore suggest the methods presented here as possible avenues for imputing data in situations where parametric assumptions need to be relaxed.

CITED LITERATURE

Babakus, E., Ferguson, C.E.Jr., Joreskog, K.G.: The Sensitivity of Confirmatory Maximum Likelihood Factor Analysis to Violations of Measurement Scale and Distributional Assumptions. Journal of Marketing Research 24; 222-228: 1987.

Barnes, S.A. Lindborg, S.R., Seaman, J.W.Jr.: Multiple Imputation Techniques in Small Sample Clinical Trials. Statistics in Medicine 25; 233-245: 2006.

Barton, R.R., Schruben, L.W.: Uniform and Bootstrap Resampling of Empirical Distributions. In Proceedings of the 25th conference on Winter simulation, pages 503-508: 1993.

Bill-Axelson, A., Holmberg, L., Ruutu, M., Garmo, H., Stark, J.R., Busch, C., Nordling, S., Häggman, M., Andersson, S-O, Bratell, S., Spångberg, A., Palmgren, J., Steineck, G., Adami, H-O, Johansson, J-E. Radical Prostatectomy versus Watchful Waiting in Early Prostate Cancer. New Engl J Med. 364; 1708-1717: 2011.

Butler, J.S., Burkhauser, R.V., Mitchell, J.M., Pineus, T.F.: Measurement Error in Self-Reported Health Variables. The Review of Economics and Statistics 69, 644-650: 1987.

Collins, L.M., Schafer, J.L., Kam, C-M.: A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. Psychological Methods 6 330-351: 2001.

Demirtas, H.: Modeling Incomplete Longitudinal Methods. Journal of Modern Applied Statistical Methods 3; 305-321: 2004.

Demirtas, H.: Simulation Driven Inferences for Multiply Imputed Longitudinal Data sets. Statistica Neerlandica 58; 466-482: 2004.

Demirtas, H.: Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. Statistics in Medicine 24; 2345-2363: 2005.

Demirtas, H.: Practical Advice on How to Impute Continuous Data When the Ultimate Interest Centers on Dichotomized Outcomes Through Pre-Specified Thresholds. Communications in Statistics – Simulation & Computation 36; 871-889: 2007.

Demirtas, H.: On Imputing Continuous Data when the Eventual Interest Pertains to Ordinalized Outcomes via Threshold Concept. Computation Statistics & Data Analysis 52; 2261-2271: 2008.

Demirtas, H., Arguelles, L.M., Chung, H. Hedeker, D.: On the Performance of Bias–Reduction Techniques for Variance Estimation in Approximate Bayesian Bootstrap Imputation. Computation Statistics & Data Analysis 51; 4064-4068: 2007.

CITED LITERATURE (continued)

Demirtas, H, Doganay, B Simultaneous Generation of Binary and Normal Data with Specified Marginal and Association Structures. Forthcoming in Journal of Biopharmaceutical Statistics: 2012.

Demirtas, H. Freels, S.A., Yucel, R.M.: Plausibility of Multivariate Normality Assumption when Imputing Non-Gaussian Continuous Ouctomes: a Simulation Assessment. Journal of Statistical Computation and Simulation 78; 69-84: 2008.

Demirtas, H. and Hedeker, D.: Gaussianization-Based Quasi-Imputation and Expansion Strategies for Incomplete Correlated Binary Responses. Statistics in Medicine 26; 782-799: 2007.

Demirtas, H., and Hedeker, D.: Multiple Imputation Under Power Polynomials. Communications in Statistics – Simulation & Computation 37; 1682-1695: 2008.

Demirtas, H. and Hedeker, D.: Imputing Continuous Data under some non-Gaussian Distributions. Statistica Neerlandica 62; 193-205: 2008.

Demirtas, H. and Hedeker D.: A Pratical Way for Computing Approximate Lower and Upper correlation bounds. The American Statistician 65; 104-109: 2011.

Demirtas, H. and Schafer, J.L.: On the Performance of Random-Coefficient Pattern-Mixture Models for Non-Ignorable Drop-Out. Statistics in Medicine 22; 2553-2575: 2003.

Dempster, A.P., Laird, N.M., and Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. JRSSB 39; 1-38: 1977.

Drasgow, F. Polychoric and polyserial correlations. Pp. 68–74 in S. Kotz and N. Johnson, eds., The Encyclopedia of Statistics, Volume 7. Wiley, 1986

Efron, B.: Missing Data, Imputation, and the Bootstrap. JASA 89; 463-475: 1994.

Emrich, J.L., Piedmonte, M.R.: A Method for Generating High-dimensional Multivariate Binary Variates. The American Statistician 45; 302–304: 1991.

Falk, M.: A Simple Approach to the Generation of Uniformly Distributed Random Variables With Prescribed Correlations. Communications in Statistics – Simulation & Computation 28; 785-791: 1999.

Farrington, D.P., and Loeber R.: Some Benefits of Dichotomization in Psychiatric and Criminological Research. Criminal Behavior and Mental Health 10; 100-122: 2000.

CITED LITERATURE (continued)

Fleishman, A.I.: A Method for Simulating Non-Normal Distributions. Psychometrika 43; 521-532: 1978.

Glynn, R.J., Laird, N.M. and Rubin, D.B.: A Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups. JASA 88; 984-993: 1993.

Gold, M.S., and Bentler, P.M.: Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. Structural Equation Modeling 7; 319-355: 2000.

Gold, M. S., Bentler, P. M., and Kim, K. H.: Comparison of Maximum-Likelihood and Asymptotically Distribution-free Methods of Treating Incomplete Non-Normal Data. Structural Equation Modeling 10; 47-79: 2003.

Guilford, J.: Psychometric Methods. New York: McGraw–Hill Book Company, Inc., 1936.

Hedeker, D. and Gibbons, R.D.: Longitudinal Data Analysis. John Wiley & Sons, Inc., Hoboken, N.J., 2006.

Heitjan, D.F. and Rubin, D.B.: Ignorability and Coarse Data. Ann. Statist. 19; 2244-2253: 1991.

Horton, N.J., and Kleinman, K.P.: Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician 61; 79-90: 2007

Hsia, J., Rodabouach, R.J., Manson, J.E., Liu, S., Frieberg, M.S., Grattinger, W., Rosal, M.C., Cochrane, B., Lloyd-Jones, D., Roberson, J.G., and Howard, B.V. Evaluation of the American Heart Association CVD Prevention Guideline for Women. Circ Cardiovasc Qual Outcomes. 3; 128-134: 2010.

Iczkowski, K.A., Torkko, K.C., Kotnis, G.R., Wilson, R.S., Huang, W., Wheeler, T.M., Abeyta, A.M., and Lucia, M.S. Pseudolumen Size and Perimeter in Prostate Cancer: Correlation with Patient Outcome. Prostate Cancer. Article 693853; 2011.

Laird, N.M., and Ware, J.H.: Random-Effects Models for Longitudinal Data. Biometrics 38; 963-974: 1982.

Le Cessie, S., and Van Houwelingen, J.C.: Logistic Regression for Correlated Binary Data. JRSSC 43; 95-108: 1994.

Li, S.T. and Hammond, J.L.: Generation of Pseudo- Random Numbers with Specified Univariate Distributions and Correlation Coefficients. IEEE Trans. on Systems, Man, and Cybernetics 5; 557-561: 1975.

CITED LITERATURE (continued)

Little, R.J.: Regression with Missing X's: A Review. JASA 87; 1227-1237: 1992.

Little, R. J. A. and Rubin, D. B. Statistical Analysis with Missing Data, 2nd ed. Wiley, New York, 2002.

Loeb, S., Han, M. Roehl, K.A., Antenor, J.A., and Catalona, W.J. Accuracy of Prostate Weight Estimation by Digital Rectal Examination versus Transrectal Ultrasonography. Journal of Urology. 173; 63-65.

Lurie, P.M., and Goldberg, M.S.: An Approximate Method for Sampling Correlated Random Variables from Partially-Specified Distributions. Management Science 44; 203-218: 1998.

McSweeney, J.C., Cody, M., O'Sullivan, P., Elberson, K., Moser, D.K., and Garvin, B.J.. Women's Early Warning Symptoms of Acute Myocardial Infarction. Circulation. 108; 2619-2623: 2003.

Mazzucchelli, R., Barbisan, F., Tarquini, L.M., Filosa, A., Campanini, N., and Galosi, A.B. Gleason Grading of Prostate Carcinoma in Needle Biopsies vs. Radical Prostatectomy Specimens. Anal. Quant. Cytol. Histol. 27; 125-133: 2005.

Montironi, R., Cheng, L., Beltran, A.L., Editorial Comment on: Comparing the Gleason Prostate Biopsy and Gleason Prostatectomy Grading System: The Lahey Clinic Medical Center Experience and an International Meta-analysis. European Urology. 54; 371-381: 2008.

Olsson, U.: Maximum Likelihood Estimation of the Polychoric Correlation Coefficient. Psychometrika 44; 443-460: 1979.

Preisser, J.S., Lohman, K.K. and Rathouz, P.J.: Performance of Weighted Estimating Equations for Longitudinal Binary Data with Drop-outs Missing at Random. Statistics in Medicine 21; 3035-3054: 2002.

Qu, Y., Piedmonte, M.R., Medendorp, S.V.: Latent Variables in Clustered Ordinal Data. Biometrics 51; 268-275: 1995.

Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. Multiple Imputation for Statistical Disclosure Limitation. J. Off. Statist. 19; 1-16: 2003.

Revelle, W. psych: Procedures for Personality and Psychological Research. R package version 1.0-92, 2011.

Rigdon, E.E., Ferguson, C.E.Jr.: The Performance of the Polychoric Correlation Coefficient and Selected Fitting Functions in Confirmatory Factor Analysis with Ordinal Data. Journal of Marketing Research 28; 491-498: 1991.

CITED LITERATURE (continued)

Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. New York: John Wiley, 1987.

Rubin, D.B.: Multiple Imputation after 18+ Years. JASA 91; 473-489: 1996.

Rubin, D.B. and Schenker, N. Multiple Imputation in Health-care Databases: An Overview and Some Applications. Statistics in Medicine 10; 585-598: 1991.

Schafer, J. L.: Analysis of Incomplete Multivariate Data. Chapman and Hall, London, 1997.

Schafer, J. L.: Multiple Imputation: A Primer. Statistical Methods in Medical Research 8; 3-15: 1999.

Schafer, J.L: Dealing with Missing Data. Res. Lett. Inf. Math. Sci. 3; 153-160: 2002.

Schafer, J.L.: Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. Statistica Neerlandica 57; 19-35: 2003.

Schafer, J. L., and Graham, J. W.: Missing data: Our View of the State of the Art. Psychological Methods 7; 147-177: 2002.

Schafer, J.L. and Olsen M.K.: Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research 33; 545-571: 1998.

Tanner, M., and Wong, W.: The Calculation of Posterior Distributions by Data Augmentation. JASA 82; 528-550: 1987.

Stampfer, M.J., Hu, F.B., Manson, J.E., Rimm, E.B., Willett, W.C. Primary Prevention of Coronary Heart Disease in Women Through Diet and Lifestyle. New Engl J Med. 343; 16-22: 2000.

Tate, R.F.: Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation. The Annals of Mathematical Statistics 25; 603-607: 1954.

Tate, R.F.: Application of Correlation Models for Biserial Data. JASA 50; 1078-1095: 1955.

Uebersax JS. The tetrachoric and polychoric correlation coefficients. Statistical Methods for Rater Agreement web site. 2006. Available at: http://john-uebersax.com/stat/tetra.htm . Accessed 04/04/2011.

Vale, C.D. and Maurelli, V.A.: Simulating Multivariate Non-Normal Distributions. Psychometrika 48; 465-471: 1983.

CITED LITERATURE (continued)

Van Buuren, S., Boshuizen, H.C., Knook, D.L.: Multiple Imputation of Missing Blood Pressure Covariance in Surival Analysis. Statistics in Medicine 18; 681-694: 1999.

Van Buuren, S.: Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. Statistical Methods in Medical Research 16; 219-242: 2007.

Yuan, K.H., and Bentler, P.M.: Three Likelihood-based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data. Sociological Methodology 30; 165-200: 2000.

APPENDIX

## Modified Lurie-Goldberg Code

```
## function for computing the marginal eCDF values for all variables in data set x.


## function obtaining lower triangular matrix from correlation matrix
l.mat=function(mat){
 k=dim(mat)[1]
 new.mat=matrix(0,k,k)
 xmat=matrix(rep(1:k,k),k,k)
 ymat=matrix(rep(1:k,each=k),k,k)
 new.mat[ymat<xmat]=mat[ymat<xmat]
 new.mat}

## function obtaining vector of pairwise correlations from
## correlation matrix
pr.cor=function(mat){
l.vec=as.numeric(l.mat(mat))
l.vec[l.vec!=0]
}

## function extracting eCDF values from ecdf function
ecdf.extract=function(y,x){
x2=apply(x,2,ecdf)
x2[[y]](x[,y])
}
## function for computing the marginal eCDF values for all variables in data set x.

mv.ecdf=function(x){
k=dim(x)[2]
k.mat=matrix(1:k,k,1)
apply(k.mat,1,ecdf.extract, x=x)
}


## function implementing Barton and Schruben (1993) method
bart.k=function(k, test.data, u.c, u.msg){

## selecting eCDF values from imputed data corresponding to Variable k missing
## entries based on imputed data
```

```
u.imp=u.c[is.na(test.data[,k])==TRUE,k]


## selecting eCDF values corresponding to Variable k observed entries

u.nonm=u.msg[is.na(test.data[,k])==FALSE,k]

## determining number of missing and observed Variable k entries
l.ui=length(u.imp)
l.un=length(u.nonm)

## determining the position where among observed eCDF values do
## eCDF values of imputed data lie

u.imp.mat=matrix(rep(u.imp, l.un), l.ui, l.un)
u.nonm.mat=matrix(rep(u.nonm, each=l.ui), l.ui, l.un)
cond.mat=matrix(as.numeric(u.imp.mat>u.nonm.mat), l.ui, l.un)
posit=apply(cond.mat,1,sum)
posit=cbind(posit, posit+1)

## sorting the observed Variable k entries in the original data and the
## corresponding observed eCDF values

ys.nonm=sort(test.data[is.na(test.data[,k])==FALSE,k])
us.nonm=sort(u.msg[is.na(test.data[,k])==FALSE,k])

## determining values in the range of the given data, based on the  eCDF values
## for the imputed data, as described in Barton and Schruben (1993)
## by first determining conditions based on the relationship between the eCDF values of
## the original and of the imputed data.

cond.1=(1:l.ui)[u.imp>min(u.nonm)&u.imp<max(u.nonm)]
cond.2=(1:l.ui)[u.imp<min(u.nonm)]
cond.3=(1:l.ui)[u.imp>max(u.nonm)]

## calculating values involving the range of the observed data and the fraction
## involving the difference between the eCDF values of the imputed and original
## data; these values will be added to the original data, as part of the back-
## transformation technique described by Barton and Schruben (1993).

add.term=1:l.ui
add.11=(ys.nonm[posit[cond.1,2]]-ys.nonm[posit[cond.1,1]])
add.12=(u.imp[cond.1]-us.nonm[posit[cond.1,1]])
add.13=(us.nonm[posit[cond.1,2]]-us.nonm[posit[cond.1,1]])

add.term[cond.1]=(ys.nonm[posit[cond.1,2]]-ys.nonm[posit[cond.1,1]])*(u.imp[cond.1]-
us.nonm[posit[cond.1,1]])/mean(us.nonm[posit[cond.1,2]]-us.nonm[posit[cond.1,1]])
add.term[cond.2]=min(ys.nonm)-abs(min(ys.nonm))*(1-(u.imp[cond.2]/min(u.nonm)))
add.term[cond.3]=max(ys.nonm)*(u.imp[cond.3]/max(u.nonm))

## performing Barton and Schruben (1993) back-transformation.

y.imp=1:l.ui
y.imp[cond.1]=ys.nonm[posit[cond.1,1]]+add.term[cond.1]
y.imp[cond.2]=add.term[cond.2]
y.imp[cond.3]=add.term[cond.3]

y.imp

}



## code for a function taking in a bivariate data set with missing entries in the second variable
## and outputting a data set with imputed values in that variable

## function for extracting the marginal eCDF values for a variable y in data set x.

ecdf.extract=function(y,x){
x2=apply(x,2,ecdf)
x2[[y]](x[,y])
}


## function for computing the marginal eCDF values for all variables in data set x.

mv.ecdf=function(x){
k=dim(x)[2]
k.mat=matrix(1:k,k,1)
apply(k.mat,1,ecdf.extract, x=x)
}

## function for computing the CDF values of a normally distributed variable based on N(xbar,S)
## where mx and sdx are the average and standard error of the data, resp.

new.pnorm=function(x){
```

APPENDIX (continued)

```
mx=mean(x)
sdx=sd(x)
pnorm(x, mx,sdx)
}

## calling library norm.
library(norm)


## function for proposed multiple imputation method incorporating
## the Lurie-Goldberg algorithm taking the values:

## test.data = data set with missing entries in the second variable  imputed.
## r.targ = target correlation (if NULL, use correlations computed from observed data)
## mn.targ = target means (if NULL, use correlations computed from observed data)
## m.imp = the number of imputations to be specified

## maxits = maximum number of iterations performed; the algorithm stops when
## either the convergence criteria is met or the maximum number of iterations is reached.

## gtol = a value multiplied to the correlation of the original data, i.e., the
## target correlation, to establish the convergence criteria.

## j.run = indicating jth simulation

 new.LG.t=function(test.data, r.targ=NULL,mn.targ=NULL, m.imp, maxits=200,gtol, j.run){

## defining matrix to hold original and imputed values for m imputed data sets

test.data3=matrix(0,dim(test.data)[1],m.imp*dim(test.data)[2])

## defining how many imputations in one simulation should be run

for (m in 1:m.imp){


## determining number of pairwise correlations and means to be computed

k=dim(test.data)[2]
kk2=k*(k-1)/2; kk3=kk2+k
if (is.null(mn.targ)==T){c.ind=numeric(kk2)}
if (is.null(mn.targ)==F){c.ind=numeric(kk3)}
```

```
## computing Pearson correlation of original data

if (is.null(r.targ)==T){r.target=cor(test.data, use="pairwise.complete.obs", method="pearson")}
if (is.null(r.targ)==F){r.target=r.targ}

## computing eCDF for given data

u.msg=mv.ecdf(test.data)
u.msg[u.msg==1]=round(1-10^-5,20)
u.msg[u.msg==0]=round(10^-5,20)

## obtaining standard normal quantiles for the eCDF values
y.msg=qnorm(u.msg)

## setting initial iteration to 0
i=0



## stopping criteria for LG algorithm
  while (sum(c.ind) < length(c.ind) & i < maxits) {
i=i+1

## imputation procedures from R norm package.

s <- prelim.norm(y.msg)   #do preliminary manipulations
thetahat <- em.norm(s, showits=FALSE, criterion=0.00001)   #find the mle

seed=sample(100:10^7,1)  #setting random number generator seed
rngseed(seed)
theta <- da.norm(s,thetahat,steps=20,showits=FALSE)  # take 20 steps
y.c <- imp.norm(s,theta,y.msg)  #impute missing data under the MLE

## obtaining normal eCDF values from imputed data
u.c=apply(y.c,2,new.pnorm)

x1=dim(test.data)[2]
na.test=apply(is.na(test.data), 2, sum)
na.test2=(1:x1)[as.numeric(na.test)>0]

x2=matrix(na.test2,length(na.test2),1)

## applying Barton and Schruben (1993) method to back-transformed eCDF values onto
```

APPENDIX (continued)

```
## original scale of the observed data
test2=apply(x2,1,bart.k,test.data=test.data, u.c=u.c, u.msg=u.msg)


## imputing the data set
test.data2=test.data
test.data2[is.na(test.data)==T]=unlist(test2)


## calculating the Pearson correlation for the imputed data set.

## computing correlations for imputed data
r.p1=cor(test.data2, method="pearson")

## computing means for imputed data
mn.p1=apply(test.data2,2,mean)

## evaluating convergence criteria involving target correlations
if (is.null(r.targ)==T){gtol.v=gtol[1:kk2]*l.mat(r.target)[l.mat(r.target)!=0]}
if (is.null(r.targ)==F){gtol.v=gtol[1:kk2]*l.mat(r.targ)[l.mat(r.targ)!=0]}

if (is.null(r.targ)==T){
c.ind=(abs(l.mat(r.p1)[l.mat(r.p1)!=0] - l.mat(r.target)[l.mat(r.target)!=0])<abs(gtol.v))
}

if (is.null(r.targ)==F){
c.ind=(abs(l.mat(r.p1)[l.mat(r.p1)!=0] - l.mat(r.targ)[l.mat(r.targ)!=0])<abs(gtol.v))
}


## evaluating convergence criteria involving target means

if (is.null(mn.targ)==F){
c.ind1=(abs(l.mat(r.p1)[l.mat(r.p1)!=0] - l.mat(r.targ)[l.mat(r.targ)!=0])<abs(gtol.v[1:kk2]))
c.ind2=(abs(mn.p1 - mn.targ)<abs(gtol[kk2b:kk3]))
c.ind=c(c.ind1,c.ind2)

}

## printing the ith iteration, mth imputation, jth simulation run, and how may pairiwise
## correlations meet the convergence criteria
```

```
print(paste("ith iteration: ",format(i), " --*-- mth imputation: ",format(m), " --*-- jth simulation run: 
",format(j.run),  " --*-- close pairs: ",format(sum(c.ind)), sep = ""))

}

## inputing m imputed data sets to be outputted into the previously defined matrix.

test.data3[,((m-1)*dim(test.data)[2]+1):(m*dim(test.data)[2])]=data.matrix(test.data2)
}

## outputs m imputed data sets.

test.data3
}
```

## Code for Inducing Correlations via Cholesky Decomposition

```
## Inducing correlations in generated bivariate data via Cholesky decomposition (Demirtas,
## personal communication).


## setting up sample size and initial rho value

n.size=100; rho=0.8


## generating a data set with missing data associated with a correlation close to the initial rho
## value
new.rho=-.5; i=0

while (abs(new.rho-rho)>.025) {

## generating an initial bivariate normal data set

library(MASS)
mydata<-mvrnorm(100, c(0,0), diag(2))

## generating an initial bivariate gamma data set

## mydata=cbind(rgamma(n.size,1,1), rgamma(n.size,1,1))

## generating an initial bivariate t-distributed data set with 3 df
## mydata=cbind(rt(n.size,3), rt(n.size,3))


i=i+1

## imposing a correlation equal to the initial rho value via Cholesky decomposition

corr<-matrix(c(1,rho, rho,1),2,2)


mydata<-mydata%*%chol(corr)

new.x1=mydata[,1]+runif(n.size)
new.x2=mydata[,2]+runif(n.size)

cor(mydata, method="pearson")

## imposing 50% MCAR in generated data set

test.mydata<-cbind(new.x1, new.x2)

test.mydata[sample(1:n.size, (n.size/2)),2]<-NA

## function to impose missingness in the second variable depends on the first variable.
## quadratically (used after correlation is induced).

mar.fxn=function(data.y, p.msg){
y1=data.y[,1]
y2=data.y[,2]
p.c=runif(100)

y2[p.msg>p.c]=NA

new.data=cbind(y1,y2)
new.data

}

## function to impose missingness in the second variable depends on the first variable
## quadratically (used after correlation is induced).

mar.fxn2q=function(data.y1){
```

APPENDIX (continued)

```
data.yq=data.y1
p.msg.e=exp(-.5+.2*data.yq[,1]+.125*data.yq[,1]^2)  ## quadratic MAR model.
p.msg.q=p.msg.e/(1+p.msg.e)

data.yq2=mar.fxn(data.yq, p.msg.q)
data.yq2
}


## imposing 40% - 50% MAR in the generated data set based on the MAR linear model

## test.mydata2=cbind(new.x1, new.x2)

## test.mydata=mar.fxn2(test.mydata2)

## imposing 40% - 50% MAR in the generated data set based on the MAR quadratic model

## test.mydata2=cbind(new.x1, new.x2)

## test.mydata=mar.fxn2q(test.mydata2)


new.rho=cor(test.mydata,  use="pairwise.complete.obs", method="pearson")[2,1]
print(i)
}


## reporting correlation associated with newly generated bivariate MCAR data
new.rho;rho
```

APPENDIX (continued)

## Code for Imputing Binary Data

```
library(norm); library(psych)


## function determining order of assignment of binary variables from underlying normally
## distributed data based on quantiles for entries where two or more values had to be imputed
new.value=function(xy, q.mix){
x=xy[1:2]; y=xy[3:4]
x.use=sample(1:4,1)
x[col.other[x.use]]=val.other[x.use]
x[col.chosen[x.use]]=as.numeric(y[col.chosen[x.use]]<q.mix[x.use])

x
}



## function to tabulate frequencies of all possible combinations of outcomes
## with binary variables

bin.p=function(x){
r=x
  n <- nrow(x)
   p <- ncol(x)
   nmis <- as.integer(apply(x, 2, sum))
   names(nmis) <- dimnames(x)[[2]]
   mdp <- as.integer((r %*% (2^((1:ncol(x)) - 1))) + 1)
   ro <- order(mdp)
   x <- matrix(x[ro, ], n, p)
   mdp <- mdp[ro]

n.mdp=table(mdp)

row.ind=as.numeric(names(n.mdp))

k=p

ind.01=c(0,1)
ir.01=NULL
for (i in 1:k){

 ir.012=rep(rep(ind.01,each=2^(i-1)),2^(k-i))
```

APPENDIX (continued)

```
ir.01=c(ir.01, ir.012)

}
ir.01.mat=matrix(ir.01,2^k,k)
ir.01.mat=cbind(ir.01.mat,0)
ir.01.mat[row.ind,(k+1)]=n.mdp
ir.01.mat
}

## function determining how many imputed values are there in an entry needing
## binary assignments
new.var=function(x){
 k=length(x)
if (k==1){nx=x}
if (k>1){
 nx=x[k]
 for (j in (k-1):1){
 nx = nx+x[j]*10^(k-j)}
}
nx
}


## function for imputing multivariate binary data for each entry
## of a multivariate binary data set
## xz = entry with binary (x) and underlying normal imputed (z) values of an entry in the data set
## y.o = data set with observed binary values
## z.o = data set with underlying normally distributed values corresponding to observed binary ##
values
## tp = frequencies of all possible combinations of outcomes with binary variables

mv.bin=function(xz, y.o=y.o, z.o=z.o,tp=tp){



## separating x and z from xz
l.xz=length(xz)

x=xz[1:(l.xz/2)]
z=xz[(l.xz/2+1):l.xz]


l.x=length(x)
```

APPENDIX (continued)

```
## if an entry has any missing values
if (sum(is.na(x))>0){

## determining observed and missing values within an entry
x.obs.ind=(1:l.x)[is.na(x)==F]
x.mis.ind=(1:l.x)[is.na(x)==T]

## if entry has one missing value
if (sum(is.na(x))==1){

## determining binary value of imputed underlying normal value based on quantiles
nx=new.var(x[x.obs.ind])
new2=apply(tp[,x.obs.ind],1,new.var)
ny=apply(y.o[,x.obs.ind],1,new.var)
z.ref=z.o[ny==nx,x.mis.ind]

## determining probabilities based on frequencies of observed combinations of binary variables
new2.tab=cbind(tp,new2)

new.prob.set=new2.tab[(new2==nx),]
prob.denom=sum(new2.tab[(new2==nx),dim(tp)[2]])
prob.num=new2.tab[(new2==nx),dim(tp)[2]][new2.tab[(new2==nx),x.mis.ind]==1]
prob=prob.num/prob.denom

## determining quantiles from probabilities based on frequencies of observed combinations of
## binary variables
q.prob=quantile(z.ref,prob, na.rm=T)

x[x.mis.ind]=as.numeric(z[x.mis.ind]<q.prob)

}

## if entry has two missing values
if (sum(is.na(x))==2){

## determining probabilities based on frequencies of observed combinations of binary variables
cnt=tp[,dim(tp)[2]]
ltp=dim(tp)[1]
prob.num=cnt[new.var(x[x.obs.ind])==apply(matrix(tp[,x.obs.ind],ltp,length(x.obs.ind)),1,new.var)]
prob=prob.num/sum(prob.num)
ntp=tp[new.var(x[x.obs.ind])==apply(matrix(tp[,x.obs.ind],ltp,length(x.obs.ind)),1,new.var),]
p11=sum(prob[apply(ntp[,x.mis.ind],1, new.var)==11])
p10=sum(prob[apply(ntp[,x.mis.ind],1, new.var)==10])
```

```
p01=sum(prob[apply(ntp[,x.mis.ind],1, new.var)==1])
p00=sum(prob[apply(ntp[,x.mis.ind],1, new.var)==0])
prob.10=p10/(p10+p00)
prob.11=p11/(p11+p01)
prob.20=p01/(p01+p00)
prob.21=p11/(p11+p10)

prob2=c(prob.10, prob.11, prob.20, prob.21)

## determining quantiles from probabilities based on frequencies of observed combinations of
## binary variables
l.y=dim(y.o)[1]; l.o=length(x.obs.ind)
z.o.sel=z.o[new.var(x[x.obs.ind])== apply(matrix(y.o[,x.obs.ind],l.y, l.o), 1,new.var),x.mis.ind]
q.mix=c(quantile(z.o.sel[,1],prob.10), quantile(z.o.sel[,1],prob.11),
quantile(z.o.sel[,2],prob.20), quantile(z.o.sel[,2],prob.21))

## determining binary value of imputed underlying normally distributed value based on quantiles
x[x.mis.ind]=new.value(c(x[x.mis.ind],z[x.mis.ind]), q.mix=q.mix)

}

## if entry has more than two missing values
if (sum(is.na(x))>2){

## determining probabilities based on frequencies of observed combinations of binary variables
lms=length(x.mis.ind)
s.ms=sample(1:lms, lms)
s.ms2=s.ms[1:2]
cnt=tp[,dim(tp)[2]]
prob=cnt/sum(cnt)
p11=sum(prob[apply(tp[,s.ms2],1, new.var)==11])
p10=sum(prob[apply(tp[,s.ms2],1, new.var)==10])
p01=sum(prob[apply(tp[,s.ms2],1, new.var)==1])
p00=sum(prob[apply(tp[,s.ms2],1, new.var)==0])
prob.10=p10/(p10+p00)
prob.11=p11/(p11+p01)
prob.20=p01/(p01+p00)
prob.21=p11/(p11+p10)

## determining quantiles from probabilities based on frequencies of observed combinations of
## binary variables
z.o.sel=z.o[,s.ms2]
y.o.sel1=y.o[,s.ms2[1]]
```

```
y.o.sel2=y.o[,s.ms2[2]]
q.10=quantile(z.o.sel[y.o.sel2==0,1],prob.10)
q.11=quantile(z.o.sel[y.o.sel2==1,1],prob.11)
q.20=quantile(z.o.sel[y.o.sel1==0,2],prob.20)
q.21=quantile(z.o.sel[y.o.sel1==1,2],prob.21)

q.mix=c(q.10, q.11, q.20, q.21)

## determining binary value of imputed underlying normally distributed value based on quantiles ## for
first two variables with missing data
new.xz=c(x[s.ms2], z[s.ms2])
new.x=new.value(new.xz, q.mix=q.mix)[1:2]
new.s=s.ms2

for (j in 3:lms){

## determining binary value of imputed underlying normally disributed value based on quantiles
## for subsequent variables with missing data
nv=new.var(new.x)
new2=apply(tp[,new.s],1,new.var)
ny=apply(y.o[,new.s],1,new.var)
z.ref=z.o[ny==nv,s.ms[j]]

cnt=tp[,dim(tp)[2]]
prob.denom=sum(cnt[new2==nv])
prob.num=cnt[new2==nv&tp[,s.ms[j]]==1]
prob=prob.num/prob.denom
q.prob=quantile(z.ref, prob, na.rm=T)
nx2=as.numeric(z[s.ms[j]]<q.prob)

new.x=c(new.x, nx2)
new.s=c(new.s,s.ms[j])

}

x[s.ms]=new.x
}
}
x
}

## setting up matrices for storing information on imputed data
n.simul=1000; m.imp=10
```

APPENDIX (continued)

```
imp.yds=matrix(0,dim(y)[1], dim(y)[2])
imp.phi12=matrix(0,n.simul,m.imp)
imp.phi13=matrix(0,n.simul,m.imp)
imp.phi23=matrix(0,n.simul,m.imp)
imp.p1=matrix(0,n.simul,m.imp)
imp.p2=matrix(0,n.simul,m.imp)
imp.p3=matrix(0,n.simul,im.imp)

for (j in 1:n.simul){

for (m in 1:m.imp){

i=0; abd.cc=rep(0,3); cc=rep(.05,3)

while (i<100 & sum(abd.cc)<3){
i=i+1

## imputing data based on joint modeling of normal data underlying binary data
## imputation procedures from R norm package.
s <- prelim.norm(nx22)   #do preliminary manipulations
thetahat <- em.norm(s, showits=FALSE)   #find the mle
rngseed(sample(10:1000,1))   #set random number generator seed
zimp <- imp.norm(s,thetahat,nx22)  #impute missing data
yz=cbind(bx3,zimp)

tp=bin.p(bx3)
y.o=bx3[apply(is.na(bx3),1,sum)==0,]
z.o=nx22[apply(is.na(nx22),1,sum)==0,]


## determining binary values from imputed underlying normally distrubed data
y.imp=t(apply(yz,1,mv.bin,y.o=y.o, z.o=z.o,tp=tp))
imp.phi12a=phi(table(y.imp[,1],y.imp[,2]),digits=4)
imp.phi13a=phi(table(y.imp[,1],y.imp[,3]),digits=4)
imp.phi23a=phi(table(y.imp[,2],y.imp[,3]),digits=4)
abd12=abs(imp.phi12a-phi12)
abd13=abs(imp.phi13a-phi13)
abd23=abs(imp.phi23a-phi23)
abd=c(abd12,abd13,abd23)
abd.cc=(abd<cc)


## printing the ith iteration, mth imputation, jth simulation run, and how many pairwise
```

APPENDIX (continued)

```
## correlations meet the convergence criteria
print(paste("ith iteration: ",format(i), " --*-- mth imputation: ",format(m),
" --*-- jth simulation run: ",format(j),
" --*-- close pairs: ",format(sum(abd.cc)), sep = ""))

}

## storing imputed data correlations and proportions involving imputed binary data
imp.yds[((j-1)*m.imp+(m-1)*dim(y)[2]+1):((j-1)*m.imp+(m-1)*dim(y)[2]+dim(y)[2])
imp.phi12[j,m]=imp.phi12a
imp.phi13[j,m]=imp.phi13a
imp.phi23[j,m]=imp.phi23a
imp.p1[j,m]=mean(y.imp[,1])
imp.p2[j,m]=mean(y.imp[,2])
imp.p3[j,m]=mean(y.imp[,3])

}
}
```

## Code for Imputing Mixed Data

```
## function for assigning binary values based on quantiles of underlying normally distributed
## values.
bind.fxn=function(x){
l.x=length(x)
l.x2=l.x/2
y=x[1:l.x2]
z=x[(l.x2+1):l.x]
p0=1-mean(y, na.rm=T)
q.p0=quantile(z,p0, na.rm=T)
z2=rep(NA,l.x2); z.o=as.numeric(na.omit(z))
ind.0=as.numeric(na.omit((1:l.x2)[y==0]))
ind.1=as.numeric(na.omit((1:l.x2)[y==1]))
z2[ind.0]=z.o[z.o<q.p0]
z2[ind.1]=z.o[z.o>q.p0]
z2
}

## function for imputing mixed data

mv.mix.fxn=function(y,n,b, n2.init, cc, maxits=100, m.imp, j.simul){
```

```
## determining conditional probabilities in binary variables that
## will be used to compute quantiles of the underlying normal variables.
y1=y[,1:n]
y2=y[,((n+1):(n+b))]
## determining entry positions of missing values
m.ind=apply(is.na(y2),1,sum)
m.ind2=(1:dim(y2)[1])[m.ind>0]
y2.m=y2[m.ind2,]
t.y2=table(y2[,1],y2[,2])
p.y2=c(t.y2[1]/sum(t.y2[1,]),t.y2[2]/sum(t.y2[2,]),
t.y2[1]/sum(t.y2[,1]), t.y2/sum(t.y2[,2]))


## computing eCDF values for continuous data
e1a=mv.ecdf(y1)

e1a[e1a==1]=round(1-10^-5,20)
e1a[e1a==0]=round(10^-5,20)

## obtaining N(0,1) values based on computed eCDF values
n1=qnorm(e1a)

## reordering values of underlying normal variables to correspond  to the binary variables
n2=apply(n2.init,2,bind.fxn)
n2.1=n2[,1]; n2.2=n2[,2]
q.y2=c(quantile(n2.2[y2[,1]==0], p.y2[1], na.rm=T),
quantile(n2.2[y2[,1]==1], p.y2[2], na.rm=T),
quantile(n2.1[y2[,2]==0], p.y2, na.rm=T),
quantile(n2.1[y2[,2]==1], p.y2[4], na.rm=T))

k=dim(y)[2]; n=dim(y)[1]

## setting up matrices to store imputed data, their pairwise correlations,
## means of imputed continuous variables, and proportions of imputed binary variables
imp.yds=matrix(0,n,k*m*j)
imp.cor2=matrix(0,j.simul,60)
imp.m1=matrix(0,j.simul,10)
imp.m2=matrix(0,j.simul,10)
imp.p3=matrix(0,j.simul,10)
imp.p4=matrix(0,j.simul,10)
```

APPENDIX (continued)

```
 for (j in 1:j.simul){
 for (m in 1:m.imp){

## setting conditions for re-iteration

abd.cc=numeric(6); i=0
  while (sum(abd.cc)<6 & i < maxits) {

i=i+1

## imputing normal data underlying original continuous and binary variables
## using the R norm package
z.n=cbind(n1,n2)
s <- prelim.norm(as.matrix(z.n))   #do preliminary manipulations
thetahat <- em.norm(s, showits=FALSE)   #find the mle
rngseed(sample(10:1000,1))   #set random number generator seed
z.imp <- imp.norm(s,thetahat,as.matrix(z.n))  #impute missing data
pz.imp=pnorm(z.imp)


## back-transforming imputed normal data onto the scale of the original continuous
## data via the Barton and Schruben (1993) method
x1=dim(y1)[2]
na.test=apply(is.na(y1), 2, sum)
na.test2=(1:x1)[as.numeric(na.test)>0]
x2=matrix(na.test2,length(na.test2),1)
y.imp1.add=apply(x2,1,bart.k,test.data=y1, u.c=pz.imp, u.msg=as.matrix(e1a))


## back-transforming imputed normal data to binary data via quantiles
z.imp2=z.imp[m.ind2,((n+1):(n+b))]
y2.2=y2[m.ind2,]
yz.imp2=cbind(y2.2, z.imp2)
y.imp2.add=t(apply(yz.imp2,1,bin2,q.y2=q.y2))

y.imp2=y2
y.imp2[m.ind2,]=y.imp2.add

y.imp=cbind(y.imp1, y.imp2)


## computing correlations of imputed data
imp.cor.m=cor(y.imp,use='pairwise.complete.obs')
imp.cor=l.mat(imp.cor.m)[l.mat(imp.cor.m)!=0]
abd=abs(imp.cor-init.cor)
dc=sign(imp.cor-init.cor)
abd.cc=(abd<cc)

## printing the ith iteration, mth imputation, jth simulation run, and how many pairwise
## correlations meet the convergence criteria
print(paste("ith iteration: ",format(i), " --*-- mth imputation: ",format(m),
" --*-- jth simulation run: ",format(j),
  " --*-- close pairs: ",format(sum(abd.cc)),
  " ** ",format(sign(dc[1])),
  " ** ",format(sign(dc[4])),
  " ** ",format(sign(dc[5])),
sep = ""))


 }


## storing imputed data correlations and proportions involving imputed binary data
imp.yds[((j-1)*m.imp+(m-1)*dim(y)[2]+1):((j-1)*m.imp+(m-1)*dim(y)[2]+dim(y)[2])
ik.dim=length(imp.cor)
imp.cor2[j,(((m-1)*k.dim+1):(k.dim*m))]=imp.cor
imp.m1[j,m]=mean(y.imp[,1])
imp.m2[j,m]=mean(y.imp[,2])
imp.p3[j,m]=mean(y.imp[,3])
imp.p4[j,m]=mean(y.imp[,4])

 }

 }

## outputting results
print(list(imp.yds, imp.cor2, imp.m1, imp.m2, imp.p3,imp.p4))

}
```

VITA

NAME: Irene B. Helenowski

EDUCATION: B.A., Molecular and Cellular Biology, Northwestern University, Evanston, Illinois, 1998

M.S., Statistics, University of Wisconsin at Madison, Madison, Wisconsin, 2001

Ph.D., Biostatistics, University of Illinois at Chicago, Chicago, Illinois, 2011

PROFESSIONAL MEMBERSHIPS: American Statistical Association

ABSTRACTS: Helenowski, I.B., Jovanovic, B.D., Chatterton, R.T., Geiger, A.S., and Gann, P.H.: Comparison of Parametric and Non-Parametric Methods for Examining the Reproducibility of Breast Fluid Biomarkers. Proceedings of the American Statistical Association, Biometrics Section (CD-ROM), paper 53, 2003.

Helenowski, I.B., Kuzel, T.M., Parameswaran, A., and Jovanovic, B.D.: Determing Cutpoints when Dichotomizing Data from the Northwestern University SPORE in Prostate Cancer. Proceedings of the American Statistical Association, Biometrics Section (CD-ROM), 2011.

PUBLICATIONS: Helenowski, I.B., Vonesh, E.F., Demirtas, H., Rademaker, A.W., Ananthanarayanan, V., Gann, P.H., and Jovanovic, B.D.: Defining reproducibility statistics as a function of the spatial covariance structures in biomarker studies. International Journal of Biostatistics. 7: 1-23, 2011.