Statistics and Its Interface Volume 6 (2013) 399–412

A semi-parametric approach for imputing mixed data

Irene B. Helenowski* and Hakan Demirtas

In this work, we present a semi-parametric method for imputing mixed data which allows us to relax assumptions of the general location model. This approach involves transforming continuous and binary variables to normally distributed data, imputing the data via joint modeling under the normality assumption, and back-transforming the data to their original scale. Transformation and back-transformation of the data comprise the nonparametric portion, and multiple imputation under the normality assumption constitutes the parametric portion of our method. Simulations involving generated mixed data with binary variables and with continuous variables following normal, t, Gamma, and mixture Gamma distributions and real data applications indicate promising results, leading us to recommend our approach as a possible avenue for imputing mixed data by semi-parametric means.

1. INTRODUCTION

Conventional approaches to imputed mixed data involve joint modeling with components of the normal model for imputing continuous data and of the saturated multinomial or loglinear model for imputing data of the binary or categorical variables (Schafer, 1997) [23]. These approaches, however, rely on assumptions pertaining to the general location model. Here, we propose a new method associated with transforming mixed data consisting of continuous and binary variables to normally distributed data, imputing these data under the normality assumption via joint modeling, and back-transforming the data onto their original scales. These transformations are implemented for continuous variables using principles of the Lurie and Goldberg (1998) [18] algorithm for generation of multivariate continuous data and empirical cumulative distribution function (eCDF) computation. The method given in Barton and Schruben (1993) [1] is involved in back-transformation of eCDF values to the scale of the original data at the final step for imputation of continuous variables. For binary variables, transformation and back-transformations of the data are associated with principles from Emrich and Piedmonte (1991) [14] and Demirtas and Doganay (2012) [9] for binary and mixed data generation, respectively. The backround involving the Lurie and Goldberg (1998) [18] algorithm is discussed in Section 3, eCDF computations are discussed in Section 4, and methods presented in Emrich and Piedmonte (1991) [14] and Demirtas and Doganay (2012) [9] are discussed in Section 5. Our new method is proposed in Section 6 and applications of the method to simulated data and real data examples are included in Sections 7 and 8, respectively. We conclude by recommending our novel approach for imputing mixed data when assumptions regarding the general location model are to be relaxed.

*Corresponding author.

2. MULTIPLE IMPUTATION

Multiple imputation, an MCMC (Markov chain Monte Carlo) technique where missing data are replaced with plausible values from a predictive distribution, has increasingly become an attractive option for handling data missing under the MCAR (Missing Completely at Random), MAR (Missing at Random), and MNAR (Missing Not at Random) mechanisms. Under the MCAR mechanism, the probability of missingness does not depend on the missing or observed data, while under the MAR mechanism, the probability may depend on the observed data. Under the MNAR mechanism, missingness depends on the missing data and may or may not depend on the observed data (Rubin, 1987 [22]; Schafer and Olsen, 1998 [25]; Little and Rubin, 2002 [17]; Demirtas and Schafer, 2003 [12]; Demirtas, 2004 [5]; Demirtas, 2004 [6]; Demirtas, 2005 [7]).

Schafer (1997) [23] introduces joint modeling as one approach for imputing data. This technique involved the Expectation-Maximization (EM) and Data Augmentation (DA) algorithms. The EM algorithm is first implemented to provide good starting values for and insight of the convergence behavior of the DA algorithm (Demirtas, 2007 [8]; Demirtas et al., 2008 [13]). The DA algorithm is comprised of two steps: the imputation step, or I-step, given in (1), where values are drawn from a distribution based on the observed data and model parameters, θ, and the posterior step, or P-step, given in (2), where parameters are updated using a distribution based on the observed and imputed data.

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

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)}) \tag{2}$$

where Y_{obs} are the observed data, Y_{mis} are the missing data, and are the imputed data.

Table 1. Simulation results for imputing bivariate mixed data involving a $N(5,1)$ distribution (Convergence constant used $= 0.01$)

Parameter	True value	AE	SB	RMSE	CR	AW
Results for MCAR case						
δ	−0.7014	−0.7011	14.3375	0.0014	93.9152	0.0901
μ_1	4.9226	4.9410				
p_2	0.4760	0.4807				
δ	−0.3799	−0.3800	3.6626	0.0014	95.7150	0.1513
μ_1	5.0317	5.0259				
p_2	0.5520	0.5494				
δ	0.4232	0.4239	41.0072	0.0015	95.5342	0.1451
μ_1	5.0481	5.0498				
p_2	0.5320	0.4987				
δ	0.7164	0.7161	14.8059	0.0014	93.0934	0.0864
μ_1	5.1496	5.1531				
p_2	0.4800	0.4917				
Results for MAR case						
δ	−0.8197	−0.8197	4.5403	0.0004	97.3606	0.0579
μ_1	5.0550	5.0917				
p_2	0.4920	0.5078				
δ	−0.2215	−0.2209	28.8064	0.0016	95.7762	0.1682
μ_1	4.9960	4.9855				
p_2	0.5360	0.5214				
δ	0.2881	0.2880	5.8686	0.0014	95.9709	0.1621
μ_1	5.0150	5.0253				
p_2	0.5040	0.4906				
δ	0.7388	0.7392	44.1313	0.0008	96.7930	0.0800
μ_1	4.9950	4.9860				
p_2	0.4440	0.4410				

The parameters updated in this algorithm are given in the general location model (Schafer, 1997 [23]; Olsen and Schafer, 2001 [21]). We define this model using p $W_1, \ldots, W_p$ categorical or binary variables and q $Z_1, \ldots, Z_q$ continuous variables comprising a data set of dimension $n \times (p+q)$. We then present the two components of the model:

$$w|\pi \sim M(n, \pi) \tag{3}$$

$$z_i|u_i = E_d, \mu_d, \Sigma \sim N(\mu_d, \Sigma) \tag{4}$$

for $d = 1, \ldots, D$, where π is the probability that an entry occurs in the d^{th} cell and μ_d and Σ are the mean vector and variance-covariance matrix associated with the distribution of the continuous variables in that d^{th} cell. Here, D is the number of cells corresponding to possible combination levels among categorical or binary variables, w is a value from a categorical or binary variable following a multinomial distribution, and z_i is a value from a continuous variables depending on a mean μ_d, $q \times q$ matrix Σ involving $u_i = E_d$, a D-vector of with 1 at position d and 0 elsewhere for $i = 1, \ldots, n$.

For example, with three binary variables, there would be $D = 2^3 = 8$ cells. Thus, the general location model can be defined in terms of a set of parameters given in (5).

$$\theta = (\pi, \mu, \Sigma) \tag{5}$$

The unrestricted model defined above is suitable when the sample size is larger than D. When some cells are associated with sparse data or zero counts, however, the means for that cell are omitted from the likelihood, and therefore the maximum likelihood estimate is no longer unique. The problem can be remedied by adding constraints to the model. For example, restrictions can be imposed on categorical variables via the loglinear model:

$$\log(\pi) = M\lambda \tag{6}$$

where π is the probability vector associated with the categorical variables as before, M is a user-specified matrix, and λ is a vector whose first element is used to scale π to sum to one. Similarly, we can impose restrictions on continuous variables. To proceed, we first consider the subset of continuous variables in relation to the multivariate regression model:

Table 2. Simulation results for imputing bivariate mixed data involving a t_3 distribution (Convergence constant used = 0.01)

Parameter	True value	AE	SB	RMSE	CR	AW
			Results for MCAR case			
δ	−0.6465	−0.6470	30.1371	0.0016	93.5712	0.1031
μ_1	−0.0827	−0.0621				
p_2	0.4880	0.5060				
δ	−0.3761	−0.3756	26.0926	0.0015	95.7219	0.1519
μ_1	−0.1206	−0.11274				
p_2	0.4560	0.4553				
δ	0.3272	0.3273	6.6861	0.0014	96.1203	0.1578
μ_1	−0.1285	−0.1251				
p_2	0.4920	0.4898				
δ	0.6500	0.6504	24.0721	0.0014	94.1362	0.1022
μ_1	−0.0446	−0.0966				
p_2	0.4760	0.4539				
			Results for MAR case			
δ	−0.6875	−0.6874	6.2143	0.0015	93.3700	0.0936
μ_1	−0.0035	−0.0582				
p_2	0.5120	0.5116				
δ	−0.2637	−0.2643	32.7540	0.0015	95.9580	0.1643
μ_1	0.1118	0.1516				
p_2	0.5080	0.4998				
δ	0.2266	0.2264	12.3985	0.0015	95.9969	0.1676
μ_1	−0.0623	−0.0886				
p_2	0.4440	0.4296				
δ	0.7027	0.7027	0.2119	0.0005	97.6648	0.0892
μ_1	0.0113	−0.0751				
p_2	0.5220	0.5344				

$$Z = U\mu + \varepsilon \tag{7}$$

for U being a $n \times D$ design matrix, μ, a vector of means, ε, the error vector, and Z, the matrix of continuous variables, $Z_1, \ldots, Z_q$, given W, the data subset of categorical variables. The mean vector μ associated with the continuous data can be constrained using:

$$\mu = A\beta \tag{8}$$

for some vector β and A being a constant matrix of dimension $D \times r$. With the restrictions, the regression model is then defined as:

$$Z = X\beta + \varepsilon \tag{9}$$

where $X = UA$ and β is associated with a reduced set of regression parameters. The coefficients of the model in equation (9) can therefore be estimable even with zero counts in the contingency table because estimation involves the rank $UA = r < D$, instead of Rank $U = D$, removing cells with zero counts from the computation.

3. LURIE-GOLDBERG ALGORITHM

Lurie and Goldberg (1998) [18] developed a method for generating multivariate continuous data using information from marginal distributions and pairwise correlation without requiring information from the joint distribution, which is often unknown. The goal of their algorithm is to minimize D^*:

$$D^* = \frac{1}{2}\sum_{i=2}^{k}\sum_{j=1}^{k-1}(r_{ij}^* - r_{ij})^2 \tag{10}$$

with r_{ij}^* and r_{ij} as the target correlation matrix and correlation matrix associated with the generated data, respectively. The algorithm involves multiplying a multivariate matrix $\mathbf{X} \sim \mathrm{N}(0, \mathrm{I})$ to the transpose of the lower triangular matrix $\mathbf{L}$ obtained via Cholesky decomposition from the correlation matrix of interest, $\mathbf{R}$, to obtain:

$$\mathbf{Y} = \mathbf{X}\mathbf{L}^{\mathbf{T}} \tag{11}$$

Probability distribution function (PDF) values based on the standard normal distribution are then calculated for $\mathbf{Y}$, as given in (12), where Φ is the cumulative distribution function of the standard normal distribution.

Table 3. Simulation results for imputing bivariate mixed data involving a Gamma$(1,1)$ distribution (Convergence constant used $= 0.01$)

Parameter	True value	AE	SB	RMSE	CR	AW
Results for MCAR case						
δ	−0.6493	−0.6502	46.5629	0.0016	93.9631	0.1023
μ_1	0.9801	0.9848				
p_2	0.5400	0.5346				
δ	−0.3946	−0.3945	8.6858	0.0014	95.8202	0.1493
μ_1	0.9122	0.9021				
p_2	0.4280	0.4455				
δ	0.4023	0.4026	17.1405	0.0015	95.5386	0.1482
μ_1	0.9315	0.9286				
p_2	0.5000	0.5222				
δ	0.6472	0.6467	27.1875	0.0015	93.9777	0.1031
μ_1	1.0071	0.9672				
p_2	0.5000	0.4890				
Results for MAR case						
δ	−0.7125	−0.7126	38.7908	0.0002	99.1663	0.0867
μ_1	1.1523	1.1028				
p_2	0.4500	0.4929				
δ	−0.4550	−0.4552	13.4534	0.0015	95.2761	0.1403
μ_1	1.0690	1.0800				
p_2	0.5040	0.5018				
δ	0.2568	0.2573	25.5506	0.0015	95.9114	0.1650
μ_1	0.9583	0.9996				
p_2	0.5300	0.5283				
δ	0.7103	0.7103	3.2183	0.0013	93.5253	0.0878
μ_1	1.0280	1.0221				
p_2	0.4720	0.4618				

$$\mathbf{U} = \Phi(\mathbf{Y}) \tag{12}$$

The inverse function of the marginal distribution is next applied to each variable Y_j, such that:

$$v_{ij} = F_j^{-1}(u_{ij}) \tag{13}$$

where the entries v_{ij} comprise the elements of the generated matrix $\mathbf{V}$. The correlation matrix $\mathbf{R}^*$, associated with $\mathbf{V}$ can then be computed, i.e.:

$$\mathbf{R}^* = \text{cor}(\mathbf{V}) \tag{14}$$

and compared to the original correlation matrix, $\mathbf{R}$. The algorithm is re-iterated until the quantity defined in (10) is less than some constant c determined by the desired level of accuracy. The criteria in (10) are designed so as to obtain the pairwise associations between variables, along with the marginal distributions desired for each variable.

4. EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION (ECDF)

In addition to principles of the Lurie and Goldberg (1998) [18] algorithm, the portion of our method for imputing continuous data also involved computation of empirical cumulative distribution function (eCDF) values. The purpose of employing eCDF computation is to relax parametric assumptions related to the distribution of the data. We define an eCDF value given in (15) for a real value x as the proportion of values in a random variable less than or equal to x, i.e.:

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \tag{15}$$

Barton and Schruben (1993) [1] present a method where eCDF values can be mapped to the scale of the original data via:

$$\begin{aligned} &F^{-1}\left(u_{C_{(i)j}}\right) = y_{(i)j} + I_{ij}, \\ &I_{ij} = (y_{(i+1)j} - y_{(i)j})\frac{u_{C_{(i)j}} - F(y_{(i)j})}{F(y_{(i+1)j}) - F(y_{(i)j})} \end{aligned} \tag{16}$$

where $F(y_{(i)j})$ and $F(y_{(i+1)j})$ are the two original eCDF values between which the new eCDF value $u_{C_{(i)j}}$ lies and $y_{(i)j}$ and $y_{(i+1)j}$ are the original data points corresponding to the two original eCDF values.

Table 4. Imputation results for trivariate data with all variables having missing entries (2 continuous variables, 1 binary variable). Order of Correlations: (Y_1, Y_2), (Y_1, Y_3), (Y_2, Y_3); *Order of Means:* Y_1, Y_2, Y_3

TRUE Correlation	Imputed Correlation	SB	RMSE	CR	AW	TRUE Means	Imputed Means	Convergence 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
$.75*\text{Gamma}(5,1) + .25*\text{Gamma}(1,1)$ 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
$\text{N}(5,1)$ results under MCR mechanism								
−0.5528	−0.5531	7.6218	0.0036	94.4052	0.2783	5.0103	5.0141	0.0250
0.4769	0.4771	5.0039	0.0039	94.2776	0.3091	4.9001	4.8883	0.0250
−0.4936	−0.4916	47.2996	0.0038	94.8941	0.3029	0.5700	0.5507	0.0225
t_3 results under MAR mechanism								
−0.4290	−0.4298	15.3363	0.0041	94.0607	0.3262	0.0494	0.0080	0.0275
0.3018	0.2996	42.0112	0.0043	94.6053	0.3628	0.4922	0.5322	0.0275
−0.4294	−0.4303	21.6933	0.0034	95.2342	0.3249	0.4000	0.4241	0.0275
$.75*\text{Gamma}(5,1) + .25*\text{Gamma}(1,1)$ results under MAR mechanism								
−0.3188	−0.3205	39.8442	0.0038	95.3493	0.3567	1.0176	0.9910	0.0250
0.4395	0.4388	19.4776	0.0031	95.7115	0.3216	4.1294	4.1585	0.0225
−0.4988	−0.5000	32.0891	0.0032	95.4791	0.2992	0.5600	0.5520	0.0225

5. BINARY AND MIXED DATA GENERATION

Emrich and Piedmonte (1991) [14] and Demirtas and Doganay (2012) [9] discuss multivariate generation of binary and mixed data from multivariate normally distributed values, respectively. In introducing these methods, we first present some pairwise correlation coefficients, such as the phi, tetrachoric, and point-biserial correlations. The phi correlation, given as:

$$\frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{10}+n_{11})(n_{00}+n_{01})+(n_{00}+n_{10})(n_{01}+n_{11})}} \tag{17}$$

for $n_{jk} = \text{I}(Y_1 = j, Y_2 = k)$, is a special case of the Pearson correlation measuring the association between two binary variables. The phi correlation is related to the tetrachoric correlation via:

$$\Phi\left[z(p_j), z(p_k), \rho_{jk}\right] = \delta_{jk}(p_j q_j p_k q_k)^{1/2} + p_j p_k \tag{18}$$

where the tetrachoric correlation measures the association of two normally distributed variables underlying the binary variables. Here, ρ_{jk} is the tetrachoric correlation, δ_{jk} is phi correlation coefficient, Φ is the bivariate normal CDF with mean 0, standard deviation 1, and correlation ρ_{jk}, p_j and p_k are the proportion parameters for Y_j and Y_k, respectively, $q_j = 1 - p_j$ and $q_k = 1 - p_k$. Lastly, $z(p_j)$ and $z(p_k)$ are the quantile functions of the standard normal distribution for p_j and p_k, respectively.

Emrich and Piedmonte (1991) [14] indicate that the multivariate normally distributed data generated using a correlation matrix with the pairwise tetrachoric correlations can then be dichotomized by quantiles based on proportions corresponding to the desired binary data. The third correlation we discuss here, the point-biserial correlation, assesses the relationship between a binary variable and a continuous variable. It is related to the Pearson correlation measuring the association between two continuous variables by:

$$\delta_{Y_1 Y_{2D}} = \left(\frac{h}{\sqrt{p(1-p)}}\right)\rho_{Y_1 Y_2}, \qquad p = \Pr(Y_{2D} = 1) \tag{19}$$

where Y_2 is a normally distributed variable underlying the binary variable Y_{2D} and h is the ordinate of the normal curve at the point defined for the binary split of Y_{2D} (Demirtas and Doganay, 2012) [9].

Table 5. Imputation results for trivariate data with all variables having missing entries (1 continuous variable, 2 binary variables). Order of Correlations: (Y_1, Y_2), (Y_1, Y_3), (Y_2, Y_3); *Order of Means:* Y_1, Y_2, Y_3

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
			N(5, 1) results under MAR mechanism					
−0.2811	−0.2804	13.5379	0.0040	94.9069	0.3668	5.1270	5.0899	0.0275
0.4799	0.4803	8.7371	0.0038	94.2611	0.3082	0.4900	0.4927	0.0275
−0.1908	−0.1904	9.4503	0.0034	95.9627	0.3820	0.5400	0.5210	0.0275
			t_3 results under MAR mechanism					
−0.3066	−0.3081	35.3747	0.0036	95.6176	0.3596	−0.1576	−0.1073	0.02325
0.5322	0.5303	45.9748	0.0035	94.9354	0.2876	0.4500	0.4967	0.02325
−0.2891	−0.2899	19.1129	0.0035	95.6721	0.3638	0.4400	0.4325	0.02325
			Gamma(5, 1) results under MAR mechanism					
−0.3180	−0.3162	40.0162	0.0041	95.0151	0.3582	4.7573	4.7396	0.0275
0.2696	0.2706	21.8347	0.0038	95.3798	0.3684	0.5100	0.5223	0.0275
−0.2932	−0.2908	47.5888	0.0046	94.5414	0.3647	0.3400	0.3838	0.0275

Demirtas and Doganay (2012) [9] extend the method of Emrich and Piedmonte (1991) [14] to generate multivate mixed data. They use principles from Emrich and Piedmonte (1991) [14] to generate binary variables. Namely, they first generate a multivariate normally distributed data set using a correlation matrix with tetrachoric pairwise correlations as defined in (18) corresponding to pairs of two binary variables, point-biserial pairwise correlations, as given in (19), corresponding to pairs of a binary and a normally distributed variable, and pairwise correlations corresponding to pairs with two normally distributed variables. After generating the multivariate normally distributed data, Demirtas and Doganay (2012) [9] then dichotomize variables designated as binary and rescale variables designated as normally distributed variables using equation (20).

$$Z_l = a_l Y_l + b_l, \quad l = 1, \ldots, q \tag{20}$$

where a_l and b_l correspond to the scale and location parameters of the desired distribution of variable Z_l.

6. NEW METHOD

We introduce our method by discussing the transformation of continuous variables and binary variables to normally distributed variables, separately. With the continuous variables, we first compute the eCDF values for the observed data and obtain corresponding normally distributed values using the inverse function of the standard normal distribution:

$$\mathrm{Y}^* = \Phi^{-1}(\mathrm{U}) \tag{21}$$

With the binary variables, we first compute the pairwise phi coefficients as given in (17) and apply equation (18) to these coefficients in order to obtain the pairwise tetrachoric correlations. We next determine if the correlation matrix is positive semi-definite. If the matrix is not positive semi-definite, we can find the positive semi-definite matrix "closest" to this matrix (Higham, 2002) [16]. We use the matrix with pairwise tetrachoric correlations to generate multivariate normally distributed data and introduce the same fraction of missing information in the multivariate normally distributed data as is present in the original binary data.

Multiple imputation via joint modeling under the normality assumption is then applied to a data set with these normally distributed values corresponding to the continuous variables and the binary variables combined. The

Table 6. Imputation results for 4-variable data with Y_2 having missing entries (2 continuous variables, 2 binary variables)

MCAR Case with missing values in Y_2							
			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.z	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)	
	0.0623	0.1988	0.5200	0.4700		4.8727	
MAR Case with missing values in Y_4							
			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_3 results				
Pairs	TRUE Value	Imputed Value	SB	RMSE	CR.z	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	
MAR Case with missing values in Y_4							
			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	

marginal PDF values of the imputed data based on the marginal distribution of the imputed data are then obtained, i.e.:

$$Y_j^{*imp} \sim N(\mu_j^{*imp}, \sigma_j^{*imp}) \tag{22}$$

for each variable Y_j following the distribution $N(\mu_j^{*imp}, \sigma_j^{*imp})$.

We next apply the equation given in (16) in order to map the imputed values pertaining to the continuous variable onto the scale of the original data. The imputed values

Table 6. (Continued)

MAR Case with missing values in Y_2							
			N(5, 1) results				
Pairs*	TRUE Value	Imputed Value	SB	RMSE	CR	AW	Convergence Constant
(Y_1, Y_2)	−0.2588	−0.2595	14.2597	0.0040	95.0299	0.3711	0.0275
(Y_2, Y_3)	0.2444	0.2431	26.7741	0.0039	95.1759	0.3741	0.0275
(Y_2, Y_4)	−0.2393	−0.2402	17.8786	0.0038	95.1718	0.3749	0.0275
	True Means (Y_1, Y_2, Y_3, Y_4)					Imputed Mean (Y_4)	
	5.0051	5.0110	0.4400	0.4700		4.9711	
			t_3 results				
Pairs	TRUE Value	Imputed Value	SB	RMSE	CR.z	AW	Convergence Constant
(Y_1, Y_2)	−0.3283	−0.3296	31.7107	0.0035	95.6598	0.3543	0.0250
(Y_2, Y_3)	0.2936	0.2944	18.3900	0.0035	95.5753	0.3630	0.0250
(Y_2, Y_4)	−0.2501	−0.2482	45.3306	0.0036	95.8699	0.3723	0.0250
	True Means (Y_1, Y_2, Y_3, Y_4)					Imputed Mean (Y_2)	
	0.0666	−0.5757	0.5500	0.5300		−0.5750	
			Gamma(5, 1) results				
Pairs	TRUE Value	Imputed Value	SB	RMSE	CR	AW	Convergence Constant
(Y_1, Y_4)	−0.3062	−0.3054	17.3540	0.0035	95.5676	0.3604	0.0258
(Y_2, Y_4)	0.4598	0.4590	16.8437	0.0038	94.6741	0.3154	0.0258
(Y_3, Y_4)	−0.3820	−0.3838	38.9096	0.0040	94.9327	0.3399	0.0258
	True Means (Y_1, Y_2, Y_3, Y_4)					Imputed Mean (Y_4)	
	4.7025	4.9320	0.4800	0.5300		4.8578	

pertaining to binary variables are lastly dichotomized by quantiles corresponding to probabilities computed for the binary data by:

$$\begin{aligned} &\Pr(Y_k = y_k | Y_1 = y_1, Y_2 = y_2, \ldots, Y_{k-1} = y_{k-1}, R), \\ &R = \{R_1, R_2, \ldots, R_{K-1}, R_K), \\ &y_k = 0, 1; \quad k = 1, \ldots, K \end{aligned} \tag{23}$$

where $R_1, \ldots, R_k$ are the missingness indicators for $Y_1, \ldots, Y_k$. The probabilities defined in (23) are computed by the equation below.

$$\begin{aligned} &\sum I(Y_k = 1 | Y_1, Y_2, \ldots, Y_{k-1}, R_1, R_2, \ldots, R_k) \\ &= \sum \frac{I(Y_k = 1, Y_1, Y_2, \ldots, Y_{k-1}, R_1, R_2, \ldots, R_k)}{I(Y_1, Y_2, \ldots, Y_{k-1}, R_1, R_2, \ldots, R_k)} \end{aligned} \tag{24}$$

The reason behind this computation is to determine the probability of the binary value for the missing entry within a subset of the data set given what other binary variables are observed.

The imputed continuous variables and the imputed binary variables are then again combined. Next, the pairwise correlation between imputed continuous variables, the pairwise phi correlations between imputed binary variables and the point-biserial correlations between imputed continuous and binary variables via equation (19) are calculated and compared to the respective biserial, phi, and point-biserial pairwise correlations obtained from the original data via the following equation:

$$\left| \delta_{jk} - \delta_{jk}^{imp} \right| < c_{jk} \tag{25}$$

where δ_{jk}^{imp} and δ_{jk} are the pairwise correlations obtained from the imputed and original data, respectively and c_{jk} is some constant chosen to achieve standardized bias values <50% and coverage rates >90% for each pairwise correlation between variables Y_j and Y_k. The algorithm is reiterated until (25) is satisfied for all pairwise correlations. Note that the selection of c_{ij} is data-specific, as it depends on the incomplete data under consideration.

We summarize the steps of our algorithm with the following diagram (Figure 1), where we note the steps at which that the continuous and binary variables are separated before transformation to normally distributed values, recombined for multiple imputation, separated for back-transformation onto the original scales, and then recombined again to create the final data set.

7. SIMULATION STUDY

We examine our method for imputing mixed data using several different bivariate and multivariate examples. Simulation studies involved generating data under MCAR and MAR mechanism with pairwise correlations ranging from −0.75 to 0.75 and including continuous variables which followed the normal, t, Gamma, or mixed Gamma distributions. Such settings were chosen to examine the applica-

Table 7. Characteristics and imputation results including assessment measures for pairwise correlations involving Prostate SPORE bivariate mixed data. Biopsy cores positive for Cancer is the continuous with 49 missing values is involved in each case. Convergence constant used = 0.01

Binary Variable Involved	No. Missing	Parameter	Original Value	AE	SB	RMSE	CR	AW
Seminal vesicle		δ	0.0909	0.0906	20.0823	0.0015	96.0800	0.1753
(No. missing = 1)	49	μ_1	2.9678	2.9828				
	1	p_2	0.0452	0.0463				
Marginal Nodes		δ	0.0960	0.0968	2.7382	0.0014	96.2520	0.1751
(No. missing = 0)	49	μ_1	2.9678	2.9936				
	0	p_2	0.1700					
Peripheral Nerves		δ	0.2710	0.2710	0.0806	0.0015	95.7429	0.1711
(No. missing = 11)	49	μ_1	2.9678	2.9700				
	11	p_2	0.6614	0.6593				

Table 8. Variables in 100 men in the Prostate SPORE database used in the imputation application to multivariate mixed data and estimates and average estimates and assessment measures from the imputation application to multivariate mixed data from the Prostate SPORE database

Variable	Label				Number (Percent) Missing	
Y_1	Percent Cancer in Prostate Gland				0 (10.0%)	
Y_2	Percent Biopsy Cores Positive for Prostate Cancer				19 (19.0%)	
Y_3	Marginal nodes (positive vs. negative)				0 (0.0%)	
Pairs	Original correlation	Imputed correlation	SB	RMSE	CR	AW
(Y_1, Y_2)	0.3366	0.3362	8.6253	0.0035	95.3717	0.353
(Y_2, Y_3)	0.264	0.2651	22.7556	0.004	95.0477	0.3699

Y_{cont}

1. $\rightarrow U_{con}$
2. $\rightarrow Y^*_{cont}$

Y^{*imp}_{cont}

6. $\rightarrow U^{imp}_{cont}$
7. $\rightarrow Y^{imp}_{cont}$

Y_{bin}

3. $\rightarrow Y^*_{bin}$

Y^{*imp}_{bin}

8. $\rightarrow Y^{imp}_{bin}$

4. $Y^* = (Y^*_{cont}, Y^*_{bin})$
5. $\rightarrow Y^{*imp}$
9. $Y^{imp} = (Y^{imp}_{cont}, Y^{imp}_{bin})$

Figure 1. Diagram indicating the steps for the new procedure of imputing mixed data, including how continuous and binary variables are separated, transformed to normally distributed values, combined, imputed, separated again, back-transformed onto the scale of the original data and finally recombined.

tion of our method to data missing under either mechanism and involving pairwise correlations of different magnitudes. Furthermore, distributions associated with continuous variables were chosen to show that our method can be applied to data with continuous variables following any distribution, whether symmetric (normal), symmetric but heavy-tailed (t), or skewed (Gamma). Bivariate data sets include 500 entries and multivariate data sets include 100 entries. Here, the number of entries refers to the number of observations and the number of missing values in each variable. Thus, for example, in bivariate cases, there are two variables and each variable contains a number of observations and a number of missing values which sum to 500. Each bivariate data set included one binary variable and one continuous variable. In both bivariate and multivariate cases, binary and continuous variables were generated separately; these variables were then combined into one data set and pairwise correlations were introduced. In multivariate examples, we generated data sets with three variables (having two continuous variables and one binary variable or one continuous variable and two binary variables) and four variables (with two continuous and two binary variables). All continuous variables involved in one multivariate data set follow the same distribution.

Under the MCAR mechanism, 25% of entries were deleted in the bivariate, trivariate, and 4-variable cases. To generate missing data in the second variable in bivariate cases under the MAR mechanism, missingness was induced in the binary variable dependent on the continuous variable via equation (26), where $R_{Y_2} = 0$ indicates that Y_2 is missing.

$$\log\left(\frac{p_2}{1-p_2}\right) = -2.75 + 0.25Y_1, \qquad p_2 = \Pr(R_{Y_2} = 0) \tag{26}$$

Missing values in the trivariate and 4-variable data sets were introduced in the second variable under the MAR mechanism using equation (27).

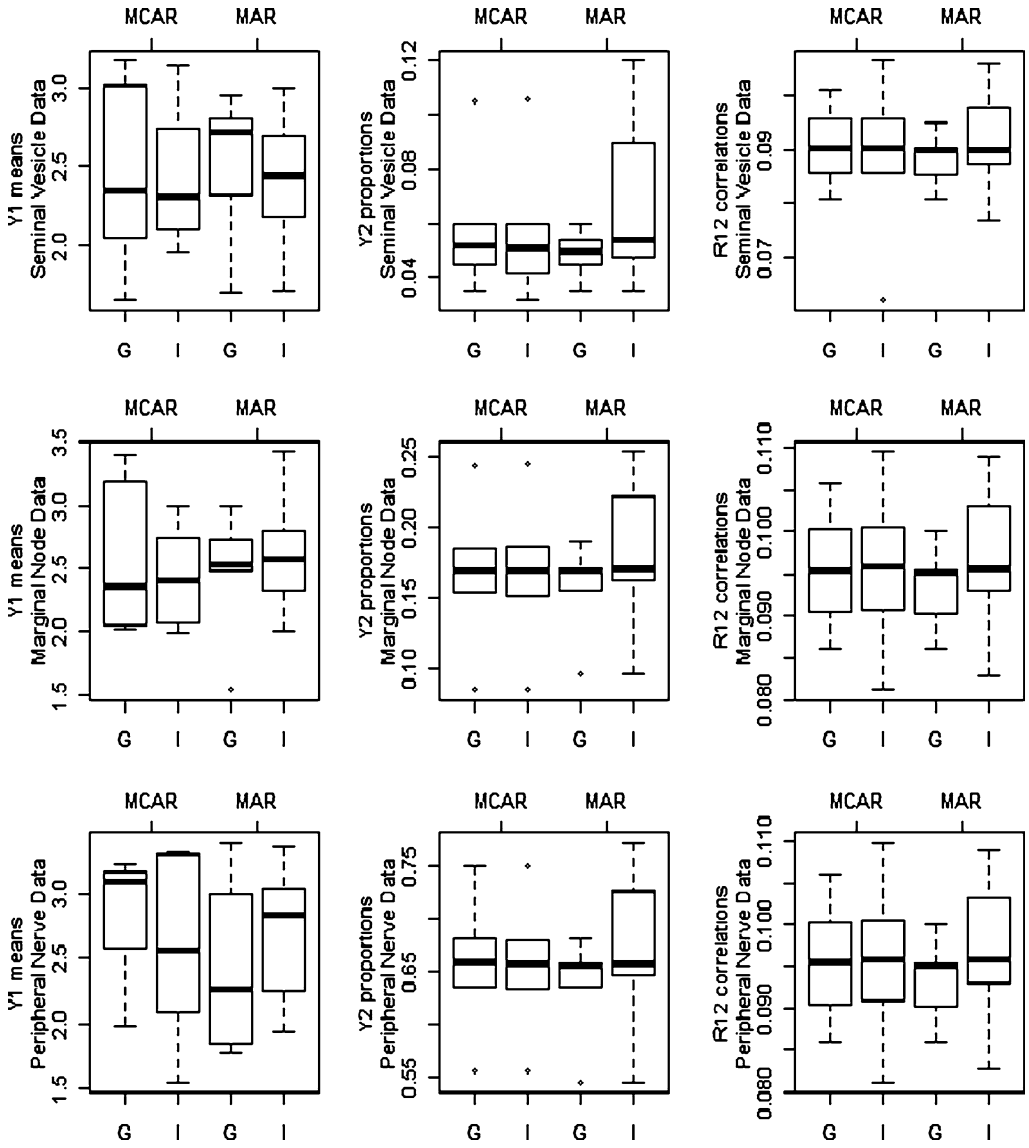

Figure 2. Boxplots of Y_1 means, Y_2 proportions, and pairwise correlations between Y_1 and Y_2 obtained from generated (G) and imputed (I) bivariate data sets resembling real data involving variables for number of biopsy cores positive for cancer (Y_1) and presence or absence of cancer in seminal vesicles, marginal nodes, or peripheral nerves (Y_2). Missing values were introduced under the MCAR mechanism in both variables or under the MAR mechanism in the binary variable and under the MCAR mechanism in the continuous variable.

$$\log\left(\frac{p_2}{1-p_2}\right) = -0.5 + 0.00025Y_3, \qquad p_2 = \Pr(R_{Y_2} = 0) \tag{27}$$

such that the probability of missingness depended on the third variable, Y_3. Missing values were then introduced in Y_1 and Y_3 under the MCAR mechanism by randomly deleting 25% of the entries in each of those variables.

We assessed the performance of our method via assessment measures for the pairwise correlations which included average estimate (AE), standardized bias (SB), the root mean square error (RMSE), the coverage rate (CR), and average width (AW) of the confidence intervals. The average estimate (AE) is the average of all parameter estimates obtained from all imputed data sets across all simulations. Standardized bias (SB), an accuracy measure, given in (28), is defined as the absolute difference between the true parameter and parameter estimate obtained from the simulations divided by the standard deviation obtained from the simulations; satisfactory SB values should be less than 50%. A standardized bias value of 50% indicates that the discrepancy between the estimate and true value is 1/2 the magnitude of the standard deviation, corresponding to 1/8 th of a typical confidence interval which is considered accept-

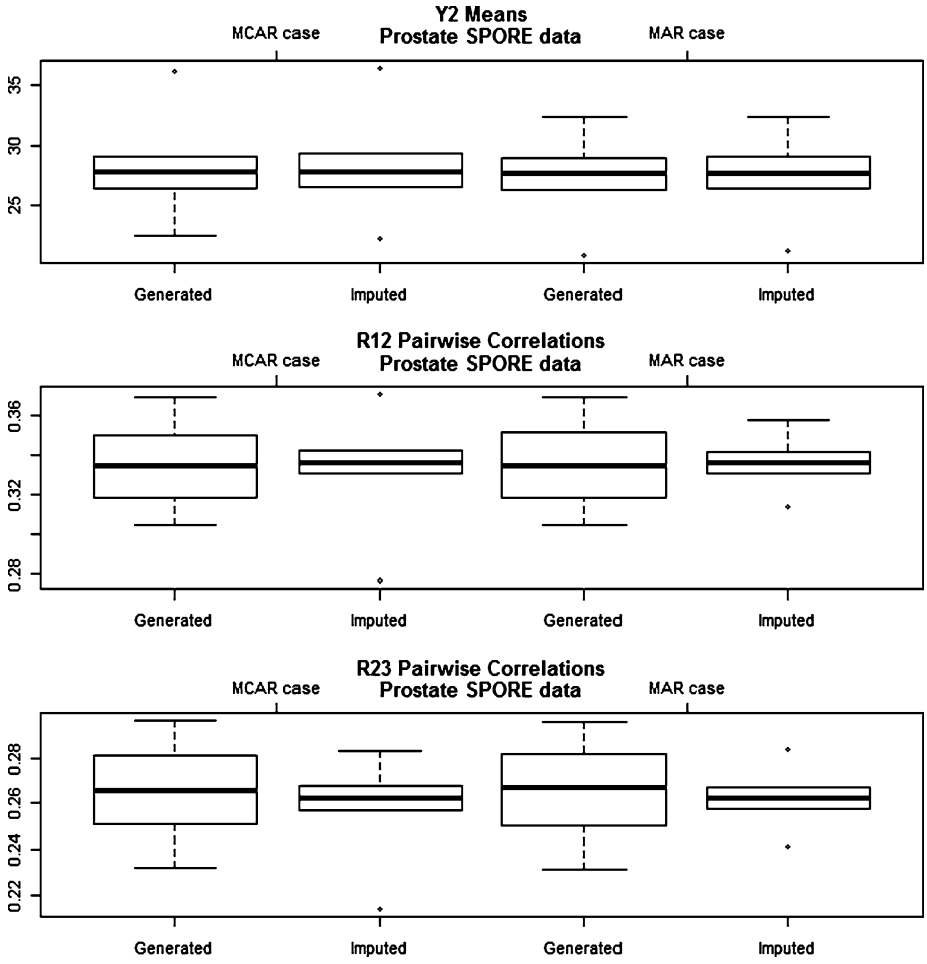

Figure 3. Boxplots of means and pairwise correlations involving percent biopsy cores positive for prostate cancer (Y_2) obtained from generated and imputed data sets for the multivariate real data examples where missing data were introduced in Y_2 via the MCAR or MAR mechanisms.

able. The root mean square error (RMSE), given in (29), is used to evaluate precision and accuracy with small values being preferable. The coverage rate (CR), again used to evaluate precision and accuracy, is defined as the percentage of times that the confidence interval of parameter estimate obtained from imputed data encompasses the true parameter, where rates >90% indicate sufficient coverage. Lastly, average width (AW), a precision measure, involves the average difference between the lower and upper bound of the confidence limits for the parameter estimate across all imputations (Collins et al., 2001 [4]; Schafer and Graham, 2002 [24]; Demirtas and Hedeker, 2007 [19]; Demirtas and Hedeker, 2008 [11]; Demirtas et al., 2008 [13]).

$$100 \times \left| \frac{E(\theta - \widehat{\theta})}{SE(\widehat{\theta})} \right| \tag{28}$$

$$\sqrt{E_\theta(\widehat{\theta} - \theta)^2} \tag{29}$$

Results indicate that our method performs adequately in both bivariate (Tables 1, 2, and 3) and multivariate cases (Tables 4, 5, and 6) as shown by AE values comparable to true parameters, SB estimates <50%, small RMSE values, CR values >90%, and AW values comparable to the 95% confidence interval widths of the true parameters for pairwise correlations.

8. REAL DATA EXAMPLE

Our real data example includes variables coming from the Prostate SPORE database. The Specialized Program of Research Excellence in Prostate Cancer (SPORE Grant #: P50 Ca 090386), established in 2001, is a collaborative effort between Northwestern University, the University of Chicago, and Northshore University Health System facilities aimed to develop new approaches for prostate cancer prevention, diagnosis, and treatment. The database contains demographic, clinical, and pathological information from 3,452

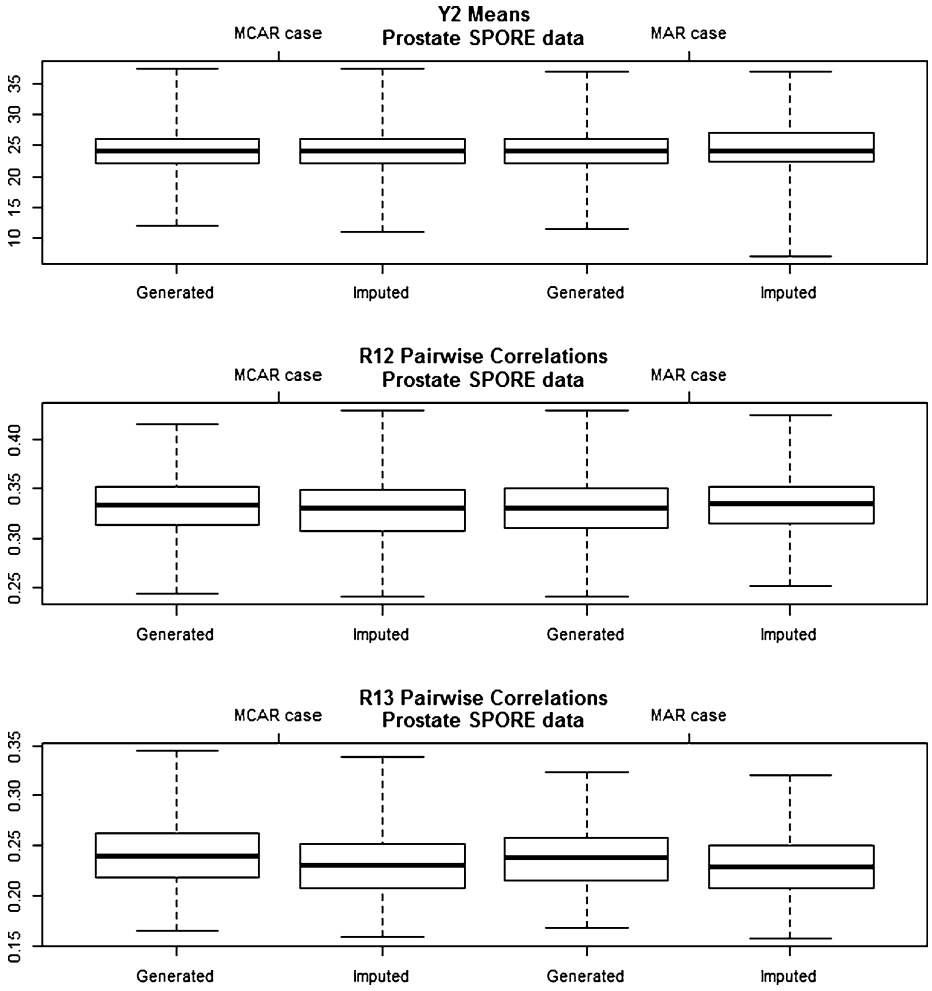

Figure 4. Boxplots of means and pairwise correlations involving percent biopsy cores positive for prostate cancer (Y_2) obtained from generated and imputed data sets for the multivariate real data examples where missing data were introduced in Y_2 via the MCAR or MAR mechanisms. Ranges for pairwise correlations from generated data and imputed data are comparable when the convergence criteria pertaining to those correlations are relaxed.

men as of 2011. Original estimates are obtained from the incomplete data. Previous analyses of these data were conducted to determine confounding factors which could contribute to bias and none were found.

Tables 7 and 8 give the descriptions of data sets used in our bivariate and multivariate examples. Here, the number of biopsies staining positive for cancer, the percentage of biopsies staining positive for cancer, and the percentage of cancer in the removed prostate gland are continuous variables, and presence of cancer in seminal vesicles, marginal nodes, or peripheral nerves are binary variables. Furthermore, the first two variables mentioned above involve data collected from needle biopsies, while the other variables were obtained from radical prostatectomy. The association between data from biopsies and from radical prostatectomy is of interest to investigators with respect to prostate cancer diagnosis as biopsy is a preferable alternative to radical prostatectomy (Mazzuchelli et al., 2005 [19]; Montironi et al., 2008 [20]; Bill-Axelson et al., 2011 [2]). With biopsy data, both number and percent positive staining have important biological implications in prostate cancer research (Gancarczyk et al., 2003 [15]; Bittner et al., 2010 [3]). Therefore, investigators often wish to examine the relationship between prognostic factors and both these variables.

Results in Tables 7 and 8 again indicate promise in the application of our method via AE values comparable to original estimates, SB estimates <50%, small RMSE values, CR estimates >90%, and AW values comparable to the 95% confidence interval widths of original estimates for the pairwise correlations.

We further generated 1,000 data sets with characteristics similar to the real data examples. The same fraction of missing information was introduced in these generated data sets as found in the original data. This missingess was in-

troduced under the MCAR mechanism via random deletion or under the MAR mechanism using equations:

$$\log\left(\frac{p_1}{1-p_1}\right) = -1.125 + 0.0005Y_2, \quad p_1 = \Pr(R_{Y_1} = 0) \tag{30}$$

in the bivariate cases and:

$$\log\left(\frac{p_2}{1-p_2}\right) = -2.0 - 0.0005Y_3, \quad p_2 = \Pr(R_{Y_2} = 0) \tag{31}$$

in the multivariate cases where one can see that missingness in a continuous variable depended on a binary variable in each case. We applied our method to each data set involving 10 imputations. Boxplots in Figures 2 and 3 indicate that means and proportions from generated and imputed data sets are acceptably comparable. Pairwise correlations for imputed data are associated with slightly narrower ranges as our multiple imputation algorithm is designed to converge when the pairwise correlations from the imputed data are sufficiently close to those of original pairwise correlations. When the convergence criteria dependent on the pairwise correlations are relaxed, i.e., when a larger c_{ij} was chosen, the ranges of the estimates from the generated and from the imputed data appear more comparable (Figure 4).

9. CONCLUSION

In this work, we present a semi-parametric approach for imputing mixed data which adopts principles from the Lurie and Goldberg (1998) [18] algorithm for generating multivariate continuous data, eCDF computation, and the Demirtas and Doganay (2012) [9] algorithm for generating mixed data, which in turn includes principles of the Emrich and Piedmonte (1991) [14] algorithm for generating binary data. Namely, principles found in Lurie and Goldberg (1998) [18] and of eCDF computation were involved in the imputation of continuous variables while imputation of binary variables was associated with principles of generating binary and mixed data as discussed in Emrich and Piedmonte (1991) [14] and Demirtas and Doganay (2012) [9]. In the latter case, pairwise phi correlations measuring the associations of binary variables were used to derive pairwise tetrachoric correlations involved in generating the multivariate normally distributed data to be imputed. eCDF computation was incorporated into our algorithm, to relax parametric assumptions on the distribution of the data. eCDF, by construction, is invariant to changes in scale and location.

This imputation technique differs from the random number generation methods in that the imputed normally distributed variables designated as continuous are mapped onto the scale of the original data via nonparametric means as found in the approach given in Barton and Schruben (1993) [1]. Simulation studies and real data applications of our method led to promising results. Therefore, we propose this approach as a possible avenue for multiple imputation of mixed data when assumptions of the general location model need to be relaxed.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Donald Hedeker, Sally Freels, Hua Yun Chen, and Borko Jovanovic for their input and recommendations in developing this approach and preparing this work. This work has been supported by the Northwestern University SPORE in Prostate Cancer (Grant #: P50 Ca 090386).

Received 8 October 2012

REFERENCES

[1] Barton, R. R. and Schruben, L. W. (1993). Uniform and Bootstrap Resampling of Empirical Distributions. In Proceedings of the 25th Conference on Winter Simulation (1993), pp. 503–508.

[2] Bill-Axelson, A., Holmberg, L., Ruutu, M., Garmo, H., Stark, J. R., Busch, C., Nordling, S., Haggman, M., Andersson, S-O., Bratell, S., Spangberg, A., Palmgren, J., Steineck, G., Adami, H-O., and Johansson, J-E. (2011). Radical Prostatectomy Versus Watchful Waiting in Early Prostate Cancer. *New Engl. J. Med.* **364** 1708–1717.

[3] Bittner, N., Merric, G. S., Galbreath, R. W., Bulter, W. M., Adamovich, E., and Wallner, K. E. (2010). Greater Biopsy Core Number Is Associated with Improved Biochemical Control in Patients Treated with Permanent Prostate Brachytherapy. *Int. J. Radiation Oncology Biol. Phys.* **78** 1104–1110.

[4] Collins, L. M., Schafer, J. L., and Kam, C-M. (2001). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods* **6** 330–351.

[5] Demirtas, H. (2004). Modeling Incomplete Longitudinal Data. *Journal of Modern Applied Statistical Methods* **3** 305–321.

[6] Demirtas, H. (2004). Simulation Driven Inferences for Multiply Imputed Longitudinal Data Sets. *Statistica Neerlandica* **58** 446–482. MR2113212

[7] Demirtas, H. (2005). Multiple Imputation Under Bayesianly Smoothed Pattern-Mixture Models for Non-ignorable Drop-Out. *Statistics in Medicine* **24** 2345–2363. MR2151710

[8] Demirtas, H. (2007). 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. MR2415691

[9] Demirtas, H. and Doganay, B. (2012). Simultaneous Generation of Binary and Normal Data with Specified Marginal and Association Structures. *Journal of Biopharmaceutical Statistics* **22** 223–236. MR2880620

[10] Demirtas, H. and Hedeker, D. (2007). Gaussianization-Based Quasi-Imputation and Expansion Strategies for Incomplete Correlated Binary Responses. *Statistics in Medicine* **26** 782–799. MR2339174

[11] Demirtas, H. and Hedeker, D. (2008). Multiple Imputation Under Power Polynomials. *Communications in Statistics – Simulation & Computation* **37** 1682–1695. MR2542425

[12] Demirtas, H. and Schafer, J. L. (2003). On the Performance of Random-Coefficient Pattern-Mixture Models for Non-ignorable Drop-Out. *Statistics in Medicine* **22** 2553–2575.

[13] Demirtas, H. Freels, S. A., and Yucel, R. M. (2008). Plausibility of Multivariate Normality Assumption when Imputing Non-Gaussian Continuous Ouctomes: A Simulation Assessment. *Journal of Statistical Computation and Simulation* **78** 69–84. MR2420089

[14] Emrich, J. L. and Piedmonte, M. R. (1991). A Method for Generating High-Dimensional Multivariate Binary Variates. *The American Statistician* **45** 302–304.

[15] Gancarczyk, K. J., Wu, H., Mcleod, D. G., Kane, C., Kusuda, L., Lance, R., Herring, J., Foley, J., Baldwin, D., Bishoff, J. T., Soderalh, D., and Moul, J. W. (2003). Using the Percentage of Biopsy Cores Positive for Cancer, Pretreatment PSA, and Highest Biopsy Gleason Sum to Predict Pathologic Stage after Radical Prostectomy: The Center for Prostate Disease Research Nonograms. *Urology* **61** 588–595.

[16] Higman, D. G. (1987). Coherent Algebras. *Linear Algebra Appl.* **93** 209–239. MR0898557

[17] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, 2nd ed.* Wiley, New York. MR1925014

[18] Lurie, P. M. and Goldberg, M. S. (1998). An Approximate Method for Sampling Correlated Random Variables from Partially-Specified Distributions. *Management Science* **44** 203–218.

[19] Mazzucchelli, R., Barbisan, F., Tarquini, L. M., Filosa, A., Campanini, N., and Galosi, A. B. (2005). Gleason Grading of Prostate Carcinoma in Needle Biopsies vs. Radical Prostatectomy Specimens. *Anal. Quant. Cytol. Histol.* **27** 125–133.

[20] Montironi, R., Cheng, L., and Beltran, A. L. (2008). 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.

[21] Olsen, M. K. and Schafer, J. L. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association* **96** 730–745. MR1946438

[22] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley, New York. MR0899519

[23] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London. MR1692799

[24] Schafer, J. L. and Graham J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods* **7** 147–177.

[25] Schafer, J. L. and Olsen M. K. (1998). Multiple Imputation for Multivariate Missing Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research* **33** 545–571.

Irene B. Helenowski
Feinberg School of Medicine
Northwestern University
Department of Preventive Medicine
680 N. Lake Shore Drive, Suite 1400
Chicago, IL 60611-4407
USA
E-mail address: i-helenowski@northwestern.edu

Hakan Demirtas
University of Illinois at Chicago
School of Public Health (MC 923)
1603 W. Taylor Street
Chicago, IL 60612-4336
USA
E-mail address: demirtas@uic.edu