Missing Data Methods when Values are Missing Together: A Simulation-Based Examination of Joint Modeling and Fully Conditional Specification Imputation Schemes

Edwin Farley[1] and Professor Roee Gutman[2]

[1] *Honors Candidate, Applied Mathematics - Computer Science, Brown University*
[2] *Honors Thesis Advisor, Associate Professor of Biostatistics, Brown University*

May 1, 2019

1 Introduction

Missing data occurs frequently in studies in medicine, the social sciences, and economics [11]. Observations can go missing due to nonresponse or attrition in subjects, or due to censorship or mistakes in recording values. Missing data itself is seldom the focus of a study, but rather a nuisance hindering progress towards a different goal. Statistical procedures often rely on complete data; in fact, an analysis that only considers observations with complete information in a data set with missing values may result in biased and inefficient estimates [12]. Imputation methods have been proposed as possible solutions to fill missing values with appropriate substitutes. We examine the performance of different imputation methods in the not-uncommon case of values missing together.

All imputation methods rely on assumptions about the mechanism that produced the missing data. Missing data mechanisms fall into three categories: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR) [11]. Missing data is described as MAR when the missingness mechanism does not depend on the missing data [11]. MCAR is a more stringent assumption, under which the missingness mechanism does not depend on the missing portion nor the observed portion of the data [11]. The validity of the MAR assumption cannot be tested, because the missing data is unobserved; however, in many cases, this assumption is plausible, or at least approximately plausible [11][12]. When data is MNAR, there are none of the independence assurances of MCAR or MAR data. The task of imputing values for missing data is more difficult in this case and development of missing data techniques for MNAR data, such as sensitivity analysis, is a topic of ongoing research [12].

The statistical literature describes different methods to implicitly and explicitly impute missing values under the MAR assumption. Implicit methods include weighting methods and methods that rely on likelihood and Bayesian modeling. Explicit statistical methods “fill in” the missing variables with one or more plausible values. Single imputation methods replace missing values with a single value, but these methods result in standard errors for the estimates that are too small. Multiple imputation methods, on the other hand, replace missing values with several plausible values [10]. While it is possible to design an efficient computational approach to imputation for a specific study, it would be time consuming to do so and its implementation would require specific expertise, so more widely-applicable methods are preferred [10]. Multiple imputation maintains generality in its simplicity, making it a popular choice to handle missing data.

Multiple imputation methods can be classified into two main groups: joint modeling and fully conditional specification. The joint modeling approach specifies a joint distribution for the missing variables from which new values for missing entries can then be sampled [9][10]. The fully conditional specification imputes values using a chain of conditional distributions, one for each variable with missing data [12]. The fully conditional specification may fail to describe a valid joint distribution with its series of univariate imputations, while the joint modeling approach has these relationships built into its imputation model [11]. On the other hand, the fully conditional specification is more flexible, and can perform adequately in cases where describing a joint distribution may be complex or the result intractable.

The aim of this paper is to compare the performance of the two multiple imputation methods in cases when there are two or more observed variables that are missing jointly. Such situations are encountered in longitudinal studies, for example, when subjects miss an appointment, leaving all planned observations for the skipped appointment missing.

We take a simulation-based approach to our investigation. Both the joint modeling and the fully conditional specification method have been implemented in packages for the ***R*** statistical programming language [5]. The **norm** package contains functions to perform joint modeling imputation under the multivariate Normal distribution [1]. The **mice** package contains functions to impute data using the fully conditional specification method [13].

Our general approach is to generate data sets and set some values to be missing or observed, varying simulation parameters such as the size of the sample, the number of variables, the proportion of rows with missing values, and the correlation between the variables. The missing values are then imputed for each data set according to both the joint modeling and the fully conditional specification method. We record performance metrics for how well each imputation method is able to maintain the characteristics of the original simulated data set and to further analyze how the performance of each imputation method may be sensitive to different parameter values. We posit that significant bias may be observed in the correlations between the covariates when using the fully conditional specification method.

2 Notation

Let $X_{i,j}$ $i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}$ represent an entry in the matrix X, where n is the number of units in the sample (rows), and m is the number of variables recorded for each unit (columns). The notation $X_{\cdot,j} = (X_{1,j}, X_{2,j}, \ldots, X_{n,j})$ refers to all the observations of the jth variable, and $X_{i,\cdot} = (X_{i,1}, X_{i,2}, \ldots, X_{i,m})$ refers to the ith unit. The data set X can be partitioned into $X = (X^{obs}, X^{mis})$ where X^{mis} is the missing part of X and X^{obs} is its observed counterpart. Similarly, the jth variable can be partitioned as $X_{\cdot,j} = (X^{obs}_{\cdot,j}, X^{mis}_{\cdot,j})$, and the ith unit as $X_{i,\cdot} = (X^{obs}_{i,\cdot}, X^{mis}_{i,\cdot})$. In addition, let R be an $n \times m$ matrix taking on values in $\{0, 1\}$, where each entry indicates whether its corresponding entry in X is missing, 0, or observed, 1.

Imputation approaches fill in X^{mis} using plausible values derived from the information in X^{obs} [9].

3 Review of Imputation Methods

3.1 Joint Modeling

The joint modeling approach to missing data imputation defines a joint model over X^{obs}, X^{mis}, and R [9]. The method imputes entries in X^{mis} based on a predefined multivariate model given the observed covariates, X^{obs}, and model parameter, θ. In practice, θ is obtained by sampling from the posterior distribution $P\left(\theta \mid X^{obs}\right)$, and imputations for X^{mis} are performed with independent draws from the posterior predictive distribution $P\left(X^{mis}|X^{obs},\theta\right)$ [9]. Formally, on iteration t the imputation procedure iterates through [9]:

$$X^{mis(t)} \overset{iid}{\sim} P\left(X^{mis}|X^{obs},\theta^{(t)}\right)$$

$$\theta^{(t+1)} \sim P\left(\theta \mid X^{obs}, X^{mis(t)}\right)$$

We assume that the rows of X are distributed according to an m-dimensional multivariate Normal distribution:

$$X \sim N(\mu, \Sigma)$$

where μ is an m-dimensional vector of means and Σ is an $m \times m$ covariance matrix.

Normality is a modeling assumption, but it has been shown that the joint modeling method with the Normal distribution maintains good operating characteristics even with “highly nonnormal” variables [11]. With this assumption, θ contains the regression parameters for the multivariate Normal regression for the variables with missing values.

A possible solution for obtaining a starting value for θ, θ^*, is maximum likelihood estimation. While there are cases in which maximum likelihood estimates can be computed directly, the EM algorithm can always be applied [9].

At iteration $t \in 1, \ldots, T$, the predictions for missing entries $X_{i,\cdot}^{mis}$, $X_{i,\cdot}^{mis(t)}$, are sampled from the corresponding posterior predictive distribution given $\theta^{(t)}$, which is multivariate Normal [9]:

$$X_{i,\cdot}^{mis(t)} \sim N\left(\theta_0^{(t)} + X_{i,\cdot}^{obs}\theta^{(t)}, \Sigma_\theta^{(t)}\right),\ i \in \{1, \ldots, n\}$$

where $\theta_0^{(t)}$ contains the intercept terms.

We can equivalently write the above as follows, by Bayes Rule:

$$X_{i,\cdot}^{mis(t)} \sim P\left(\theta_0^{(t)} + X_{i,\cdot}^{obs}\theta^{(t)}|\Sigma_\theta^{(t)}\right) \cdot P\left(\Sigma_\theta^{(t)}\right),\ i \in \{1, \ldots, n\}$$

It turns out that $P(\Sigma_\theta)$ corresponds to an inverse Wishart distribution [9]. Thus, we can sample a value for $\Sigma_\theta^{(t)}$, and use it to fully define the multivariate Normal from which we draw $X_{i,\cdot}^{mis(t)}$ [11][9].

After a complete draw of values in X^{mis}, a new θ value, $\theta^{(t+1)}$, is sampled from its posterior distribution [9]:

$$\theta^{(t+1)} \sim N\left(\theta^{*}, n^{-1}\Sigma^{(t+1)}\right)$$

$$\Sigma^{(t+1)} \sim \text{Inv-Wishart}\left(n-1, (nS)^{-1}\right)$$

where S is the sample covariance matrix. The covariance matrix $\Sigma_{\theta}^{(t+1)}$ is a transformation of $\Sigma^{(t+1)}$.

The joint modeling approach aims to preserve the properties of the joint distribution, namely, the means, variances, and covariances. By preserving these properties, any response-predictor relationship can be maintained in the completed data [11]. The joint modeling approach is implemented in ***R*** by the **norm** package [5][1].

3.2 Fully Conditional Specification

The joint modeling approach requires that a joint distribution for all of the variables with missing data in the data set be defined, but in many practical applications, defining such a distribution may be complex or intractable. A flexible approach that has been proposed to address this issue is the fully conditional specification [12]. This method imputes missing values variable-by-variable by defining an imputation model for each variable with missing data, conditioned on the remaining variables [12]. Using these conditional densities, this approach attempts to approximate a Bayesian parametric model [7][12]. Imputed values are generated by simulating draws from these conditional distributions [10]. The conditional distributions may be such that they can be written explicitly for certain problems, but, in more complex models, samples can be generated using computational techniques, such as Markov chain Monte Carlo. One full iteration in the fully conditional specification approach involves iterating over every variable-specific model to generate a new value [12]. This structure means that the fully conditional specification can define models for which a joint distribution does not exist, because it only requires the specification of univariate conditional models [13].

Formally, the fully conditional specification defines a conditional model for each variable with missing data, $P(X_{\cdot,j}|X_{\cdot,-j}, \theta)$, where θ is the model parameter and $X_{\cdot,-j}$ refers to the entire data set except for column $X_{\cdot,j}$. Assuming $P(\theta)$ is the prior distribution for θ, the posterior distribution of θ can be defined as $P(\theta|X^{obs}) \propto P(X_{\cdot,j}|X_{-j}, \theta) \cdot P(\theta)$.

The fully conditional specification proceeds by sampling from the posterior distribution through iterated sampling from conditional distributions of the form [13]:

$$P(X_{\cdot,1}|X_{\cdot,-1},\ \theta_1)$$

$$\vdots$$

$$P(X_{\cdot,m}|X_{\cdot,-m},\ \theta_m)$$

The complete sampling process is similar to a Gibbs sampler for the imputed variables, $X_{\cdot,1},\dots,X_{\cdot,m}$, and the corresponding unknown parameters by which the missing values are imputed, $\theta_1,\dots,\theta_m$. From the tth iteration estimates, $(\hat{X}_{\cdot,1}^{(t)},\dots,\hat{X}_{\cdot,m}^{(t)},\hat{\theta}_1^{(t)},\dots,\hat{\theta}_m^{(t)})$, the values for iteration $t+1$ are calculated as draws from the following chained conditional distributions [13]:

$$\begin{array}{lcl}
\hat{\theta}_1^{(t+1)} & \sim & P(\theta_1\ |X_{\cdot,1}^{obs}, X_{\cdot,2}^{(t)},\dots,X_{\cdot,m}^{(t)}) \\
\hat{X}_{i,1}^{(t+1)} & \sim & P(X_{i,1}|X_{\cdot,1}^{obs}, X_{i,2}^{(t)},\dots,X_{i,m}^{(t)},\hat{\theta}_1^{(t+1)}),\ i\in\{1,\dots,n\} \\
 & \vdots & \\
\hat{\theta}_m^{(t+1)} & \sim & P(\theta_m\ |X_{\cdot,m}^{obs}, X_1^{(t)},\dots,X_{m-1}^{(t)}) \\
\hat{X}_{\cdot,m}^{(t+1)} & \sim & P(X_1|X_{\cdot,m}^{obs}, X_1^{(t)},\dots,X_{m-1}^{(t)},\hat{\theta}_m^{(t+1)}),\ i\in\{1,\dots,n\}
\end{array}$$

where $X_{\cdot,j}^{(t)}=(X_{\cdot,j}^{obs},\hat{X}_{\cdot,j}^{(t)})$ is the variable $X_{\cdot,j}$, complete with imputed values at iteration t.

Note that in the sampling procedure, each $X_{i,j}$ in $X_{\cdot,j}^{mis}$ is sampled according only to other values in unit i and no other information from the data set except the information that enters through θ_j. For the variables that have no missing values, only the regression parameters are sampled, so the above sequence of equations can be simplified depending on the application. The fully conditional specification approach is implemented in ***R*** by the **mice** package [5][13].

4 Experimental Methods

In our simulation, the variables $X_{\cdot,1}$ and $X_{\cdot,2}$ are the only variables with missing data. Missing entries are missing jointly in these variables. Therefore, only quantities related to $X_{\cdot,1}$ and $X_{\cdot,2}$ will be affected by imputation, such as the mean values and correlation coefficients. In particular, for examining the correlation, we are only concerned with the first two columns, or first two rows, of the correlation matrix, because those are the entries that correspond to relationships involving $X_{\cdot,1}$ and $X_{\cdot,2}$.

4.1 Experimental Design

Our experiment takes the form of a full factorial design over the parameters H, C, n, m, r, and init. The parameters H and C define the correlation between variables in the generated data sets, where H is the correlation between X_1 and X_2, and C is the correlation between all other pairs of variables. The parameters n and m define the size of the generated data set, where n is the number of rows and m is the number of variables, so the number of variables without missing data is $m-2$. The parameter r defines the proportion of rows with missing data. Lastly, the parameter init specifies the initialization method for the fully conditional specification. The levels of each parameter that are considered are listed in Table 1.

Table 1: Parameter values

Parameter	Levels
H	$\{0.1, 0.5, 0.9\}$
C	$\{0.1, 0.5, 0.8\}$
n	$\{500, 1000, 2000\}$
m	$\{4, 8, 16\}$
r	$\{0.1, 0.3, 0.5\}$
init	$\{\texttt{none}, \texttt{linreg}, \texttt{full}\}$

4.2 Data Generation

For each value of H and C, a data set is generated according to a multivariate Normal distribution with mean $\vec{0}$ and correlation matrix Σ. The entries in Σ corresponding to the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$ have value H and all other off-diagonal entries are set to the value of C.

To reduce computational complexity and achieve greater precision, our simulation generates data sets with the greatest value of m and the greatest value of n. For different simulation configurations of m and n we then use the data from the first m columns and the first n rows extracted from the full generated data set.

Values are set to be missing in the full data set so that the expected proportion of missing data is r. The probability that $X_{\cdot,1}$ and $X_{\cdot,2}$ are missing for any given observation depends on the value of $X_{\cdot,3}, \ldots, X_{\cdot,m}$, making this missing data mechanism explicitly MAR. A logistic regression with parameter vector β is used to determine missingness. The probability of having missing data in row i, p_i, is as follows:

$$p_i = \frac{e^{X_i^* \cdot \beta}}{1 + e^{X_i^* \cdot \beta}}$$

where $X_i^* = (X_{i,3}, \ldots, X_{i,m})$ is the vector of observed values in row i.

The values for $\beta_1, \ldots, \beta_{m-2}$ are set randomly for each trial, and the intercept term, β_0, is found using the `optim` one-dimensional optimization method in ***R***, such that the average probability of missingness over all rows is equal to r. Then, for all i, $X_{i,1}$ and $X_{i,2}$ are set to be missing with probability p_i. This process is repeated in each replication of each configuration, so the same rows do not necessarily have missing values from one replication to the next.

For the initialization parameter, init, a value of `none` indicates that random initialization will be used; this is the default behavior in the **mice** package [13]. The `linreg` initialization creates two linear models, one for $X_{\cdot,1}$ and one for $X_{\cdot,2}$, in terms of $X_{\cdot,3}, \ldots, X_{\cdot,m}$, and passes the estimates for missing values of $X_{\cdot,1}$ and $X_{\cdot,2}$ according to the linear models as the initial values. The `full` option passes the complete data set so that the true value of each missing field is the initial value for imputation. The initialization values are calculated for each configuration and are only used with the fully conditional specification method.

4.3 Imputation and Metrics

We run 100 replications for each configuration of H, C, r, m, n, and init, and 5 full imputations are performed in each replication. The same metrics are recorded for imputations under the **mice** and **norm** imputation methods. In each replication, we record the sample means of $X_{\cdot,1}$ and $X_{\cdot,2}$ and the correlation matrices for each of the imputed data sets.

From the correlation matrices, we calculate the average correlation bias for the values that involve $X_{\cdot,1}$ or $X_{\cdot,2}$. This leaves us with $2m-3$ bias terms. We record the bias for the relationship between $X_{\cdot,1}$ and $X_{\cdot,2}$, as well as the average bias across the remaining $2m-4$ correlation values.

While we cannot pick a representative correlation out of the $2m-4$ relationships between $X_{\cdot,1}$, $X_{\cdot,2}$ and $X_{\cdot,3}, \ldots, X_{\cdot,m}$, it is meaningful to focus on the correlation between the two variables with missing data, $X_{\cdot,1}$ and $X_{\cdot,2}$. Therefore, we would like to create a confidence interval to determine whether the sample correlation after imputation differs from its true value at a statistically significant level. We turn this into a binary coverage metric by checking for each replication whether the true correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$ is included in a 95% confidence interval about the correlation estimated after imputation.

The Fisher Transformation allows for testing of hypotheses about the value of the population correlation coefficient, ρ, between two variables when applied to the sample correlation coefficient, r. Formally, the Fisher Transformation is:

$$f(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

The quantity $f(r)$ is asymptotically distributed as a Normal distribution with mean $f(\rho)$ and standard deviation $\frac{1}{\sqrt{s-3}}$, where s is the size of the sample. For our purposes, squaring this standard deviation gives us the within-imputation variance, which we will call U. The between-imputation variance, B, is the variance of the set of transformed correlation values, r_f, obtained from each of the $s = 5$ imputed data sets. Using Multiple Imputation combination rules we obtain the total variance, T [10][7]:

$$T = \left(1 + \frac{1}{5}\right)B + U$$

To construct a confidence interval for the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$ we rely on the Student's t-distribution with v degrees of freedom [10]:

$$v = (4)\left[1 + \frac{U}{T-U}\right]^2$$

Using this t-distribution we can construct a 95% confidence interval for the mean of the transformed correlation values from the 5 imputed data sets as

$$\left(\bar{r}_f - \alpha_{0.975}\sqrt{T},\ \bar{r}_f + \alpha_{0.975}\sqrt{T}\right)$$

where $\alpha_{0.975}$ is the 97.5th percentile under the standard t-distribution with v degrees of freedom. We then determine whether the true population correlation coefficient under the Fisher Transformation, $f(\rho)$, is in the confidence interval, setting the coverage metric to 1 if $f(\rho)$ is in the confidence interval and to 0 otherwise.

5 Results

Our aim is to examine how the imputation of missing values by different methods affects the relationships between variables in the imputed data sets. Preliminary analyses found that the means of the variables with missing data were not meaningfully affected by imputation under either approach. Therefore, we will focus on the correlation to quantify the way in which imputation preserves the relationships between variables.

We will focus primarily on the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$, so unless otherwise specified, a reference to correlation bias refers to this correlation relationship. We first examine the coverage to understand which parameters significantly affect the performance of the imputation methods in preserving the relationship between $X_{\cdot,1}$ and $X_{\cdot,2}$. The fully conditional specification approach tends to misrepresent the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$, which is explored more deeply in an analysis of the mean squared error of the correlation and the bias.

Changes observed in coverage, mean squared error, and bias in the results over samples with data missing together stand in contrast to the performance of the imputation methods in identical simulations in which entries were not missing together. This analysis is not included here, but it allows us to conclude that the effects on performance that we have observed, and now report, arise from the fact that the missing values are missing jointly, and depend on how the methods under consideration handle imputation in this case.

5.1 Correlation Confidence Interval Analysis

If the value of a simulation parameter or the imputation method were to have no effect on coverage, we would expect 95% of samples to have a coverage value of 1, matching the predefined nominal level. Therefore, a trend in the average coverage deviating from 0.95 for a particular imputation method would suggest the presence of a parameter-driven or method-driven effect when it comes to preserving the relationship between the jointly missing variables during imputation. Even observing a coverage rate greater than 95% has implications for performance, indicating that the method may be inefficient with high sample variance and a wide confidence interval as a result.

An ANOVA analysis of the coverage metric for both the **mice** package and the **norm** package can indicate which parameters influence the performance in this metric. Table 2 shows the Mean Squared Error (MSE) corresponding to the **mice** package and **norm** package by simulation parameters.

Table 2: ANOVA analysis for correlation coverage

Parameter	MSE (**mice**)	MSE (**norm**)
H	3.3903	0.0392
C	1.3572	0.0204
n	0.1222	0.3749
m	1.6807	0.0580
r	1.1702	0.2822

Nearly all the MSEs for **mice** are larger than those for **norm**. Most notably, H, the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$, C, the correlation between $X_{\cdot,1}$, $X_{\cdot,2}$, and the rest of the variables, and m, the number of variables in the data set are highly influential for **mice**. For the **norm** package, no parameters are highly significant, compared to **mice**. This suggests that the joint modeling approach using the **norm** package is less sensitive to different parameter configurations in terms of coverage. Furthermore, the parameters that are relatively significant for **norm** are the same parameters that are significant for both techniques when values are not missing together, and have roughly the same magnitudes as observed under joint modeling in this case.

The **mice** package has an initialization option for imputation. Table 3 shows the average coverage value by initialization method for **mice**.

Table 3: Proportion of samples with coverage by initialization (init)

Initialization	Coverage
`none`	0.9432
`linreg`	0.9404
`full`	0.9262

All of the values are less than 0.95, but the `full` initialization has the worst operating characteristics in this metric. There are two possible explanations for this behavior. It could be that starting at the true value does not produce enough variability in estimates, so even relatively accurate imputations are outside of an extremely narrow confidence interval, or that the bias is high. Table 4 shows the total variance and correlation bias by initialization.

Table 4: Total variance and average correlation bias by initialization (init)

Initialization	Total Variance	Bias
`none`	0.00199	-0.00274
`linreg`	0.00198	-0.00231
`full`	0.00185	-0.00416

The `full` initialization has the smallest variance and the largest bias. It seems that starting from a simple random value (`none`), or a relatively informed estimate (`linreg`), allows the sampling method in the fully conditional specification to arrive at values that better preserve the relationship between the missing variables than starting at the true values.

The dimensionality of the data set, m, is also an influential simulation parameter. Table 5 shows the coverage proportion for data imputed under both methods by values of m.

Table 5: Proportion of samples with coverage by m

m	**mice**	**norm**
$m = 4$	0.9440	0.9458
$m = 8$	0.9383	0.9432
$m = 16$	0.9276	0.9431

There is little deviation from 0.95 for the joint modeling approach when using the **norm** package; however, a significant drop in coverage is observed as the number of covariates increases under the fully conditional specification approach with **mice**. This could be a result of reduced variance or greater bias. The total variance and the correlation bias for the fully conditional specification method are summarized in Table 6.

Table 6: Total variance and average correlation bias by m (**mice**)

m	Total Variance	Bias
$m = 4$	0.00189	-0.00171
$m = 8$	0.00194	-0.00231
$m = 16$	0.00199	-0.00520

The variance is similar across values of m, but the bias increases with m. This reveals that the fully conditional specification approach results in larger bias as the number of covariates grows. In practice, one could omit variables for imputation, but this contradicts common findings that all available variables should be used for imputation [4][6].

A second-degree ANOVA analysis shows that H and C together are the most influential pair of parameters in explaining coverage deviations. Table 7 shows the average coverage by H, C.

Table 7: Proportion of samples with coverage by H, C

	mice			**norm**		
	$H = 0.1$	$H = 0.5$	$H = 0.9$	$H = 0.1$	$H = 0.5$	$H = 0.9$
$C = 0.1$	0.9415	0.9499	0.9431	0.9442	0.9458	0.9437
$C = 0.5$	0.9462	0.9394	0.9051	0.9406	0.9458	0.9426
$C = 0.8$	0.9479	0.9352	0.9212	0.9447	0.9451	0.9442

The **norm** package shows little deviation from the 0.95 benchmark. On the other hand, there is a clear trend in the values for the **mice** package, in which the coverage decreases as H increases, for all values of C. The decrease in coverage is from approximately 0.95 to 0.90. This shows that the fully conditional specification method is affected by the correlation structure defined by H and C.

5.2 Correlation Mean Squared Error Analysis

To compare the performance of the **mice** package and the **norm** package in terms of the correlation MSE we examine the ratio of the MSE values for the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$ from imputations under the fully conditional specification (**mice**) to the values under the joint modeling approach (**norm**). A ratio greater than 1 indicates a larger MSE with **mice** and a ratio less than 1 indicates a larger MSE with **norm**. Table 8 shows the MSE ratio by value of H and C.

Table 8: MSE ratio (**mice**/**norm**)

	$H = 0.1$	$H = 0.5$	$H = 0.9$
$C = 0.1$	1.0349	1.0224	1.0399
$C = 0.5$	1.0153	1.0932	1.4261
$C = 0.8$	1.0137	1.1616	1.4805

The ratio of MSEs is always greater than 1, and it increases with H across values of C. Thus, on average, the fully conditional specification yields larger errors than joint modeling.

The MSE can also be expressed as the sum of the variance and the squared bias. Table 9 shows the ratio of sampling variances, **mice** to **norm**, for each pair of parameter values.

Table 9: Total variance ratio (**mice**/**norm**)

	$H = 0.1$	$H = 0.5$	$H = 0.9$
$C = 0.1$	1.063202	1.038959	1.004529
$C = 0.5$	1.054043	1.046438	1.002913
$C = 0.8$	1.043272	1.053894	1.029638

All ratios are larger than 1, and decrease as H increases. This suggests that the trend that was observed in the MSE ratio is driven by increased bias.

5.3 Correlation Bias Analysis

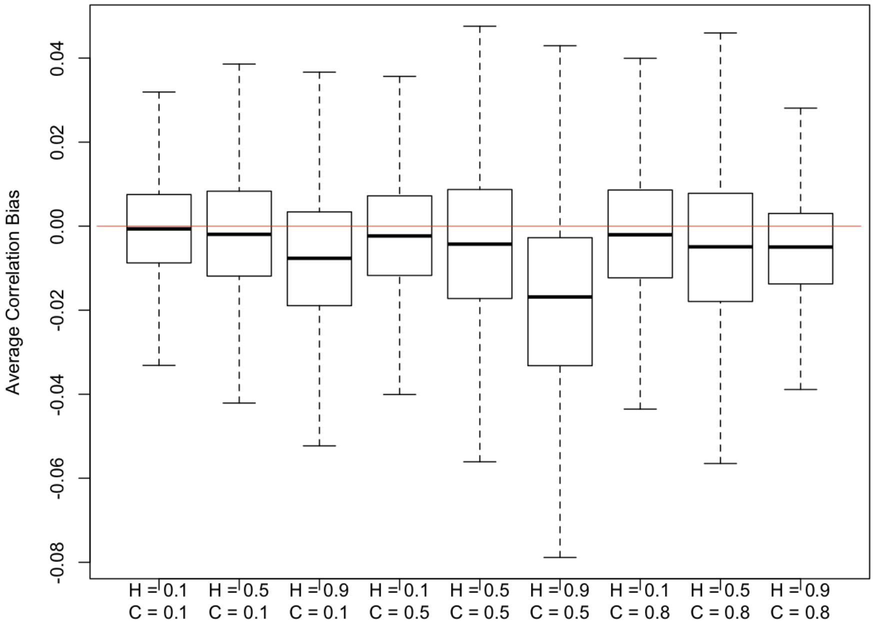
The bias terms for the joint modeling approach are at least an order of magnitude smaller than the corresponding bias under the fully conditional specification approach (Table 10). Therefore, the remainder of our analysis of the

Table 10: Correlation bias by H, C

	mice			**norm**		
	$H = 0.1$	$H = 0.5$	$H = 0.9$	$H = 0.1$	$H = 0.5$	$H = 0.9$
$C = 0.1$	-0.00196	-0.00177	-0.00113	-0.00108	-0.00090	-0.00003
$C = 0.5$	-0.00138	-0.00517	-0.00466	0.00009	-0.00190	-0.00026
$C = 0.8$	-0.00281	-0.00519	-0.00356	-0.00011	0.00024	-0.00022

correlation bias will focus on the fully conditional specification. Figure 1 depicts the average correlation bias in terms of H and C for the fully conditional specification approach.

Figure 1: Average correlation bias vs. H, C



Across values of C, the relationship between the imputed variables is consistently underestimated at higher levels of H with median bias values that are below 0.

Separating the bias of the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$, and the average of the correlation biases for the relationships between $X_{\cdot,1}$, $X_{\cdot,2}$ and the remaining variables, $X_{\cdot,3}, \ldots, X_{\cdot,m}$, can provide additional insight on the interaction between the simulation parameters and the correlation bias. We expect that the operating characteristics with respect to the relationships between $X_{\cdot,1}$, $X_{\cdot,2}$ and $X_{\cdot,3}, \ldots, X_{\cdot,m}$ would be better than the correlation between $X_{\cdot,1}$ and

$X_{\cdot,2}$, because the basis for the imputation of $X_{\cdot,1}$ and $X_{\cdot,2}$ is the set of fully observed values of $X_{\cdot,3}, \ldots, X_{\cdot,m}$.

Table 11 shows the correlation bias for the relationship between $X_{\cdot,1}$ and $X_{\cdot,2}$, and the average correlation bias over all the correlations including $X_{\cdot,1}$ or $X_{\cdot,2}$ with $X_{\cdot,3}, \ldots, X_{\cdot,m}$, in terms of H and C.

Table 11: Correlation bias by H, C: $X_{\cdot,1}$ and $X_{\cdot,2}$ (**A**), and $X_{\cdot,3}, \ldots, X_{\cdot,m}$ (**B**)

	A			**B**		
	$H = 0.1$	$H = 0.5$	$H = 0.9$	$H = 0.1$	$H = 0.5$	$H = 0.9$
$C = 0.1$	-0.00196	-0.00177	-0.00113	-0.00040	-0.00154	-0.00884
$C = 0.5$	-0.00138	-0.00517	-0.00466	-0.00194	-0.00483	-0.02103
$C = 0.8$	-0.00281	-0.00519	-0.00356	-0.00238	-0.00477	-0.00703

Generally, the correlation bias between $X_{\cdot,1}$ and $X_{\cdot,2}$ is greater than the corresponding correlation bias for the other variables; however, the trend of worse performance with higher values of H that was clear in the previous analysis is mostly observed for the average correlations between $X_{\cdot,1}$, $X_{\cdot,2}$ and $X_{\cdot,3}, \ldots, X_{\cdot,m}$ when it comes to bias. In fact, for $H = 0.9$ the alternative bias consistently exceeds that of the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$.

6 Discussion

The main difficulty with imputing values that are always missing together is that, for a single unit with missing data, there is no information from either variable to inform the imputation of the other. This challenge is especially relevant to the performance of the fully conditional specification method, in which a full multivariate model is not defined. The joint modeling approach provided valid estimates of the relationships between the variables, while the fully conditional specification resulted in large bias under certain simulation configurations.

There are certain cases where this difficulty is mitigated. For example, when $X_{\cdot,1}$ and $X_{\cdot,2}$ are weakly correlated with other variables and with each other, the fully conditional specification approach yields an average correlation bias that is close to 0. In addition, as the correlation between the variables that are missing together increases, the variability of the correlation increases, except when the correlation with other variables is also strong. When the correlation between $X_{\cdot,1}$ and $X_{\cdot,2}$ is strong, the imputation should generally be easier, because the value of $X_{\cdot,1}$ can be determined by the value of $X_{\cdot,2}$, and vice versa. Nevertheless, when examining the correlation bias, we observed that strong correlation only leads to better operating characteristics when $X_{\cdot,1}$ and $X_{\cdot,2}$ are strongly correlated with the other variables as well. When the correlation among all variables is strong, the fully conditional specification model can take advantage of these relationships and is able to provide unbiased correlation estimates [3]. These two cases stand in contrast to instances in which there is medium correlation between the variables. For example, when the correlation between the variables that are missing together is high and there is medium correlation with other variables ($C = 0.5$), the estimated correlation has larger bias and larger sampling variance.

The performance of the **mice** package under the `full` initialization leaves some unanswered questions. The `full` initialization is infeasible in practice, because the missing data is not known, but the worse performance of this method compared to the other methods is surprising, because it would appear to be the ideal initialization. Larger bias and slightly smaller sampling variance was responsible for the low coverage when using the `full` initialization, but this behavior seems to be limited to the correlation between the variables that are missing together. Biases for the correlations between the variables with missing data approximately doubled, but the average bias of the remaining correlations halved. The initialization method does not seem to influence the effect on performance of the other parameters that we have considered, because the correlation biases under the `full` initialization still follow the same general trend that has been discussed with respect to H and C. Taking a closer look at the case of the `full` initialization could reveal significant insight as to the nature of the fully conditional specification.

A few ways to expand this study would be to consider cases in which more than two variables have data missing together, or when other variables happen to have missing values as well. In addition, this study only considered continuous variables, and it would be worth expanding its scope to cover binary and discrete variables as well. These additions could serve to better mimic practical studies and provide additional guidance for investigators.

7 Conclusion

This thesis has examined the performance of the fully conditional specification and the joint modeling approaches to missing data imputation in the case when a pair of variables are always missing together. Both methods were effective in preserving the means of the variables with missing data; however, diverging trends were observed in the operating characteristics when examining the correlations of variables after imputation. Our simulations showed that while the joint modeling approach was not affected by the parameters that were examined, the performance of the fully conditional specification was affected by different design parameters when values were missing jointly. In particular, the interaction between the correlation between the variables missing together and the correlations between all other pairs of variables was a strong driver of performance.

We conclude that the fully conditional specification may struggle in this situation. The fully conditional specification imputes variables with missing data one at a time with a series of variable-by-variable models, each conditioned on the other variables. When values are missing together, the imputation of one of the variables depends on an imputed value in the other variable (or variables, potentially), exacerbating the challenge of imputation. This stands in contrast to the joint modeling approach, in which a joint model is specified to impute all of the missing values at once. The joint modeling approach is thus able to better preserve the correlations in a data set with values missing together. Values *can* be imputed with minimal negative effects with the fully conditional specification, but only under certain circumstances in which the joint modeling approach is no worse, and certainly more consistent, as an alternative.

The implications of the biased correlations created by imputation under the fully conditional specification are clear. Statistical analysis of data that contains missing values is made possible by imputation methods that complete the data; however, if in doing so, the relationships between variables are not preserved, analyses that directly examine or are sensitive to the correlation between variables may result in biased estimates.

References

[1] Ported to R by Alvaro A. Novo. Original by Joseph L. Schafer jlsstat.psu.edu. (2013). norm: Analysis of multivariate normal datasets with missing values. R package version 1.0-9.5.

[2] Little R. J. A., Rubin D. B. Statistical analysis with missing data. second edition Wiley, 2002.

[3] Liu, Jingchen, et al. "On the stationary distribution of iterative imputations." *Biometrika* 101.1 (2013): 155-173.

[4] Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 538-558.

[5] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[6] Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American statistical Association, 91(434), 473-489.

[7] Rubin, D. B. *Multiple imputation for nonresponse in surveys.* New York: John Wiley, 1987.

[8] Rubin D. B., and Joseph L. Schafer. Efficiently creating multiple imputations for incomplete multivariate normal data. 1990 Proceedings of the Statistical Computing Section, American Statistical Association 1990.

[9] Schafer, Joseph L. *Analysis of incomplete multivariate data.* Chapman and Hall/CRC, 1997.

[10] Schafer, Joseph L. "Multiple imputation: a primer." *Statistical methods in medical research* 8.1 (1999): 3-15.

[11] Schafer, Joseph L., and John W. Graham. "Missing data: our view of the state of the art." Psychological methods 7.2 (2002): 147.

[12] Van Buuren, Stef. "Multiple imputation of discrete and continuous data by fully conditional specification." *Statistical methods in medical research* 16.3 (2007): 219-242.

[13] Van Buuren, Stef and Karin Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R." *Journal of statistical software* (2010): 1-68.