Multiple Imputation for Continuous and Categorical Data: Comparing Joint and Conditional Approaches

Jonathan Kropko*
University of Virginia

Ben Goodrich
Columbia University

Andrew Gelman
Columbia University

Jennifer Hill
New York University

October 4, 2013

Abstract

We consider the relative performance of two common approaches to multiple imputation (MI): joint MI, in which the data are modeled as a sample from a joint distribution; and conditional MI, in which each variable is modeled conditionally on all the others. Implementations of joint MI are typically restricted in two ways: first, the joint distribution of the data is assumed to be multivariate normal, and second, in order to use the multivariate normal distribution, categories of discrete variables are assumed to be probabilistically constructed from continuous values. We use simulations to examine the implications of these assumptions. For each approach, we assess (1) the accuracy of the imputed values, and (2) the accuracy of coefficients and fitted values from a model fit to completed datasets. These simulations consider continuous, binary, ordinal, and unordered-categorical variables. One set of simulations

*Corresponding author: jkropko@virginia.edu. We thank Yu-sung Su, Yajuan Si, Sonia Torodova, Jingchen Liu, and Michael Malecki, and two anonymous reviewers for their comments. An earlier version of this study was presented at the Annual Meeting of the Society for Political Methodology, Chapel Hill, NC, July 20, 2012.

uses multivariate normal data and one set uses data from the 2008 American National Election Study. We implement a less restrictive approach than is typical when evaluating methods using simulations in the missing data literature: in each case, missing values are generated by carefully following the conditions necessary for missingness to be "missing at random" (MAR). We find that in these situations conditional MI is more accurate than joint multivariate normal MI whenever the data include categorical variables.

1 Introduction

Multiple imputation (MI) is an approach for handling missing values in a dataset that allows researchers to use the entirety of the observed data. Although MI has become more prevalent in political science, its use still lags far behind complete case analysis – also known as listwise deletion – which remains the default treatment for missing data in Stata, R, SAS, and SPSS. Complete case analysis (CCA) entails deleting every row in the data that has any missing values. This practice reduces the available degrees of freedom for model estimation and deletes perfectly valid data points that happen to share a row with a missing value. If a survey respondent declines to report his or her income, for example, we can no longer use this respondent's sex, age, or political preferences in our model unless we are willing to drop income from the analysis. If the observed data for the people who choose not to report their income are different from the data for the people who do state their income, then CCA can return inaccurate parameter estimates.

Every approach to MI follows the same two steps: (1) replace the missing values in the data with values that preserve the relationships expressed by the observed data; and (2) use independently drawn imputed values to create several datasets, and use the variation across these datasets to inflate model standard errors so that they reflect our uncertainty about the parametric imputation model.

In practice, however, there are many ways to implement MI, and these approaches differ greatly in the assumptions they make about the structure and distribution of the data. This study is a comparison of the two most commonly used MI algorithms: joint multivariate normal (MVN) MI and conditional MI. The phrase "joint MI" can refer to any MI algorithm in which the data are assumed to follow a known joint probability distribution with unknown parameters. Nearly all implementations of joint MI in practice, however, make the assumption that the data are MVN. Since the multivariate normal is a purely continuous distribution, any non-continuous variables are typically modeled and imputed as continuous and are then assigned to discrete values at the end of the process.[1] In contrast, conditional MI draws imputations from conditional distributions that are flexible to the type and distribution of each variable. Although both joint MVN MI and conditional MI are implemented in commonly used software, very little research provides practical, empirically tested guidance to researchers as to which approach to MI is the best option for a particular data structure.

[1] When there are only a few discrete variables, they can be combined with the multivariate normal in the general location model (Schafer 1997) but in settings such as survey data where most if not all variables are discrete, such a model does not in general have enough structure to produce reasonable imputations, hence the widespread use of MVN imputation despite its theoretical problems.

We aim to provide this guidance by simulating missing data using several different data generating processes (DGPs). We begin by simulating data that matches the assumptions made by joint MVN MI: all variables are continuous and are generated from a MVN distribution. We then take small steps to consider more general data structures: variables are discretized to be binary, or categorical with between 3 and 10 ordinal or nominal categories. Finally, we dispel the assumption of multivariate normality and consider data from the 2008 American National Election Study (ANES). In each simulation, we use a novel approach to generating missing values that conforms to Rubin's (1987) definition of missing at random (MAR) but is a more general approach than is typical in the missing data literature. We compare joint MVN MI to conditional MI as well as CCA and bootstrap draws from the observed values. We evaluate performance based on the accuracy of their imputations and on the accuracy of coefficients and predictions from a model run on imputed data. While these comparisons do not consider every theoretical approach to MI, they focus on the two strategies which dominate the use of MI in applied research.

2 Background

Our goal in this study is to consider the advantages and disadvantages of joint MVN MI and conditional MI, their assumptions, and the effect of these assumptions on the quality of the imputations and on the accuracy of parameter estimates for a model that uses the imputed data. These comparisons are important because joint MVN MI and conditional MI are currently the two dominant approaches to multiple imputation in applied research, and because very few studies have attempted to compare these methods in a systematic and empirical fashion.

The dominance of joint MVN MI and conditional MI is demonstrated by the fact that that these methods are the only two approaches to MI that are implemented in the base-packages of most widely used statistical software packages.[2] While specialized software exists for some alternative MI algorithms,[3] joint MVN MI and conditional MI are the most ubiquitous algorithms. Furthermore, several textbooks on missing data analysis place a special emphasis on joint MVN MI (Schafer 1996, Little and Rubin 2002), or on both joint MVN MI and conditional MI (van Buuren 2012), above alternative approaches.

Van Buuren (2007), Lee and Carlin (2010) and Yu, Burton, and Rivero-Arias (2007) previously compared joint MVN MI and conditional MI with mixed results: van Buuren and Yu, Burton, and Rivero-Arias find that conditional MI outperforms joint MVN MI, while Lee and Carlin find that joint MVN MI performs at least as well as conditional MI. The scope of these studies, however, is somewhat limited. Both Van Buuren and Lee and Carlin compare the ability of joint MVN MI and conditional MI to return accurate coefficient estimates from a

[2] In Stata 13, for example, joint MVN MI is implemented by the `mi impute mvn` command and conditional MI is implemented by the `mi impute chained` command (StataCorp 2013) and through the user-written commands `uvis` and `ice` (Royston 2005, 2007, 2009; van Buuren, Boshuizen, and Knook 1999). In SAS 9.0, joint MVN MI and conditional MI are implemented respectively as the `mcmc` and `reg()` options of `proc mi` (Yuan 2013). Several libraries implement joint MI in R, including `Amelia` (Honacker, King, and Blackwell 2011, 2012) and `Norm` (Schafer and Olsen 1998). Conditional MI is implemented in R by the `mice` package (van Buuren and Groothuis-Oudshoorn 2011) and by the `mi` package (Goodrich et al 2012, Su et al 2011).

[3] For example, Schafer (1997) proposes a log-linear joint distribution for datasets that only include categorical variables and a general location model to fit the joint distribution of data that contain both continuous and categorical variables, and distributes software on his webpage to implement these procedures in S-Plus.

regression model. Van Buuran uses complete cases as the standard against which to compare the two approaches, while Lee and Carlin only consider a model with a complete continuous outcome and partially observed binary or ordinal predictors. Neither study discusses the accuracy of the imputed values or of the model's fitted values. Yu, Burton, and Rivero-Arias (2007) compare the imputed values from a number of joint MVN MI and conditional MI algorithms to the true values of several continuous variables, but do not consider categorical variables or the results from a regression model. In addition, none of the software packages that implement joint MVN MI and conditional MI include any diagnostics or advice to researchers who are trying to choose between the two algorithms. Stata's (2013) documentation explicitly states that it makes "no definitive recommendation" on this decision (p. 124). In short, researchers who are trying to choose between joint MVN MI and conditional MI still lack clear guidance. We aim to provide this guidance by considering the effect of the MI algorithms on both the accuracy of regressions and of missing value imputations using partially observed continuous, binary, ordinal, and unordered-categorical variables.[4] We begin with a more detailed description of the joint MVN and conditional approaches to MI.

2.1 Joint Multivariate Normal MI

Joint MI specifies a joint distribution of the data, estimates the parameters of this joint distribution, and draws imputed values from this distribution. Most implementations of joint MI – including the ones considered in this study – use the assumption that the data follow a joint MVN distribution. If the data are distributed according to a known distribution, imputing missing values is only a simple matter of drawing from the assumed distribution. In order to be less restrictive, implementations of joint MVN MI sometimes use a version of the EM algorithm that alternates between estimating the means, variances, and covariances of the MVN distribution and drawing new imputed values (Dempster, Laird, and Rubin 1977). Given this flexibility, joint MVN MI is able to accurately impute missing values for any data configuration that resembles MVN. The two joint MVN MI algorithms considered here – `Amelia` (Honacker, King, and Blackwell 2011, 2012) and `Norm` (Schafer and Olsen 1998) – use variants of the EM algorithm.[5] If the data are not close to MVN, however, there is no reason to expect that the imputed data will be accurate.

In addition to the MVN assumption, discrete and limited variables must be treated as if they were continuous variables – since the MVN distribution is only defined for continuous dimensions – then reconstructed once the imputations are complete. One way to turn continuous imputed values into discrete values is to round them to the nearest category. For example, an algorithm that treats a binary variable as continuous will produce continuous

[4] Van Buuren (2012) argues that analyses that consider imputation accuracy are invalid because imputation algorithms that ignore the uncertainty of the imputations – such as replacing missing values with the mean of the observed values of each variable – can produce smaller RMSEs than algorithms that do account for the noise. However, the conditional and joint MVN MI algorithms considered in this paper meet Rubin's (1987) definition of proper: that is, they model the noise around the imputed values. The measures of accuracy considered here provide a fair adjudication of MI algorithms that are all proper.

[5] These two programs differ only in their approach to modeling the variance across imputed datasets. `Amelia` adapts the EM algorithm to include bootstrapping (Honacker and King 2010). `Norm` simulates noise by drawing imputed values from their individual distributions. This approach is a Markov Chain Monte Carlo simulation, and is guaranteed eventually to converge in distribution to the joint posterior distribution of μ, Σ, and the imputed values (Schafer and Olsen 1998, p. 555).

imputed values that are often within $[0, 1]$ but sometimes outside this range. After imputations are complete, imputed values within $(-\infty, .5)$ can be rounded to 0, and values within $[.5, \infty)$ can be rounded to 1. A similar technique can be applied to ordinal variables. Binary and ordinal rounding is somewhat intuitive since an imputed value of .2 is probably more similar to a real value of 0 than 1. Rounding can also be applied to unordered-categorical variables, which is less intuitive since intermediate values between two categories may not have any substantive meaning.

Rounding is a controversial topic that has gotten some attention in the MI literature. Some researchers suggest that rounding makes imputed values less accurate. Horton, Lipsitz, and Parzen (2003) consider the imputation of binary variables under joint MVN MI, and show that there exist situations in which the unaltered continuous imputations for binary variables are more accurate than the rounded imputations. Honacker, King, and Blackwell (2012), the authors of `Amelia`, suggest that users allow some non-categorical values. They note that sometimes a decimal value may represent an intermediate value between two ordinal categories (p. 16-17). Other times, however, continuous imputations simply are not reasonable values for categorical variables.

As an alternative to simple rounding, `Amelia` uses the Bernoulli, binomial, and multinomial distributions to create binary, ordinal, and unordered-categorical draws from continuous imputed values (Honacker, King, and Blackwell 2012, p. 17-18). Binary and ordinal variables are imputed as if they are continuous. These imputed values are continuous and may be greater than the numerical value of the highest category, less than the numerical value of the lowest category, or in between the numerical values of the other categories. Values which are imputed to be outside the range of the categories are rounded to the nearest category. `Amelia` then equates the continuous imputed values with probabilities. Continuous draws for binary variables are used as parameters for independent Bernoulli distributions, and an imputed value of 0 or 1 is randomly drawn from each distribution. Continuous draws for ordinal variables are scaled to be within $[0, 1]$ and are used in independent binomial distributions, where a categorical value is randomly drawn from each distribution. Unordered-categorical variables are broken into binary indicators for each category, and these binary variables are imputed together. Imputed values outside $[0, 1]$ are rounded to 0 or to 1, and the values are scaled so that they sum to 1. These values are treated as probabilities which are passed to independent multinomial distributions, and an imputed category is randomly drawn from each multinomial distribution.

Alternative approaches have also been proposed. Bernaards, Belin, and Schafer (2007) compare simple rounding and Honacker et al's probabalistic approach to a technique in which the cutoff for rounding to 0 and 1 – typically .5 – is instead estimated. Although they find support for the accuracy of this approach, it is only applicable to binary and some ordinal imputations. Demirtas (2010) proposes a rule for ordinal variables that breaks an ordinal variable into indicator variables, and selects the category whose indicator has the highest modal probability.[6]

[6]More specifically, for an ordinal variable with k categories, Demirtas defines a matrix which appends a $(k-1) \times (k-1)$ identity matrix to a $(k-1) \times 1$ row of zeroes to represent each category, then takes the vector of $k-1$ continuous imputations and calculates the Euclidean distance to every row of this matrix. The category whose row has the lowest distance is drawn as the imputation.

All of these approaches draw an equivalence between the continuous imputed values and probability. A different approach that avoids rounding from probability distributions involves drawing imputed values using predictive mean matching (PMM) with joint MVN MI. PMM is an alternative to drawing from the posterior predictive distribution (PPD) for modeling the noise in the data, and is used by hotdeck imputation techniques (Cranmer and Gill 2013). PPD and PMM are described in more detail in appendix 1, and we duplicated our analysis using PMM to model the noise for both joint MVN MI and conditional MI. These supplementary analyses are presented in appendix 2, but do not change our conclusions about the relative strengths of joint MVN MI and conditional MI.

2.2 Conditional MI

Conditional MI involves iteratively modeling the conditional distribution of each partially observed variable, given the observed and imputed values for the other variables in the data. Like joint MVN MI, conditional MI requires assumptions about the distribution of the data. Conditional MI capitalizes on the fact that the vast majority of statistical models that exist are conditional models, therefore it is easier to find models tailored to the specific features of each variable type. By focusing on the conditionals, conditional MI is more flexible than joint MVN MI, and can model a wider class of joint distributions.

The conditional MI algorithm involves two nested loops. The first loop iterates over the partially observed variables, replacing the missing values with new imputed values at each step. In order to derive new imputed values, conditional MI commonly uses an appropriate generalized linear model for each variable's conditional distribution in which the variable is fit against all the other variables. The `mi` package in R (Goodrich et al 2012, Su et al 2011) by default uses Bayesian versions of OLS, logit, and ordered logit for continuous, binary, and ordinal variables respectively, and multinomial logit for unordered-categorical variables.[7] In these models, all of the predictor variables are treated as if they are complete; the missing values are replaced with their most recent imputed values. The algorithm draws imputed values including the appropriate amount of predictive uncertainty and replaces the previously imputed values with these new values.[8] This first loop is nested inside a second loop that repeats the process until the conditional distributions appear to have converged and to be drawing from the same joint distribution of the data. Finally, several chains of this algorithm are run independently in order to assess whether they have converged to the same distribution after a given number of iterations.

Iterative draws from conditional distributions approximate a joint distribution even when the joint distribution is not known or is non-analytic. In fact, if the joint distribution is truly multivariate normal, then the joint MVN MI algorithm described in section 2.1 is equivalent to a conditional MI algorithm that uses OLS for each conditional

[7] Multinomial logit is implemented in the `nnet` package (Venables and Ripley 2002). In addition, count variables are modeled using a Bayesian quasi-poisson regression (Gelman et al 2012), interval variables are modeled with a parametric survival regression model (Therneau 2012), proportions are modeled with a beta regression (Cribari-Neto and Zeileis 2010). These models are defaults, and can be altered to alternative or to custom models by the user.

[8] Starting values for the imputed values are drawn from the empirical marginal distribution of each variable. That is, the missing values are initially replaced with randomly drawn (with replacement) values from the observed datapoints for each variable. Different draws are made for every chain, and when the chains converge, it implies that the starting values do not influence the results.

distribution. Conditional MI, however, may be a viable option even when we cannot safely make an assumption about the joint distribution.

There are some important similarities between joint MVN MI and conditional MI. First, both joint MVN MI and conditional MI conceive of imputed values as draws from the joint distribution of the data, even if they make different assumptions about the joint distribution. Second, both joint MVN MI and conditional MI require multiple imputed datasets in order to correctly model the uncertainty due to imputation. Finally, neither joint MVN MI nor conditional MI are likely to be accurate estimators of the imputed values when the data are strongly NMAR.

A major critique of conditional MI is that the algorithm resembles – but is not – a Gibbs sampler, and the conditionals are not necessarily compatible with a real joint distribution. While a Gibbs sampler is guaranteed to eventually converge to the correct posterior distribution, conditional MI is not. Li, Yu, and Rubin (2012) refer to conditional MI's algorithm as a possibly incompatible Gibbs sampler (PIGS), and demonstrate that there are situations in which a PIGS may not converge at all, or else might converge to a different distribution if the variables are imputed in a different order. Li, Yu, and Rubin do not demonstrate, however, that there exists a joint distribution for which a PIGS performs badly and a joint MVN algorithm does well. Any joint distribution has conditionals, and so long as the conditionals are correctly specified a conditional algorithm should not suffer from the problems identified by Li, Yu, and Rubin.

Our goal is to examine which style of algorithm tends to be more appropriate in general settings. Using simulations, described in sections 3 and 4, we find that conditional MI typically returns more accurate imputed values and model parameters than joint MVN MI for data that are generated from a MVN distribution and discretized, and also for commonly used political science data.

3 Simulations Based on Multivariate Normal Data

This simulation uses artificial data generated from a MVN distribution. We consider four cases: in every case, a variable is set aside as the variable "of interest"; in the first case this variable remains continuous; in the second case the variable is turned binary; in the third case the variable is turned ordinal; and in the fourth case the variable is constructed to be unordered-categorical. Since the initial distribution of the data is MVN, these simulations present the most favorable circumstances for joint MVN MI and isolate the impact of the discrete value conversions required for joint MVN MI.

We begin with complete data and impose MAR missingness onto the data. We run competing MI algorithms on the partially observed data and assess their performances using two standards: (1) the difference between the imputed values and the true values, and (2) the similarity between a regression model run on the imputed data to the same regression run on the true data. In each of the continuous, binary, ordinal, and unordered categorical cases, we generate 8 variables. We intend for three of these variables to be complete and for five variables – one of which

may be categorical – to have missing values.[9] When the variable of interest is ordinal or unordered-categorical, we repeat the simulation 8 times, increasing the number of categories from 3 to 10. For each simulation, we run 1000 iterations in which we generate MVN data with MAR missingness, run several competing MI algorithms, and assess their performances.

There are three obstacles to generating informative MVN data with MAR missingness. First, we intend to consider the family of MVN distributions rather than assume a distribution with a fixed covariance matrix. Second, we require a reasonable method for discretizing continuous variables. Third, we want to simulate missingness patterns in the data which are as general as possible while still conforming to the MAR assumption. We consider our strategy for addressing these challenges to be in itself a contribution to the missing data literature. Our method is briefly discussed here, and is described in detail in appendix 3.

In order to consider the family of MVN distributions as opposed to a single fixed distribution, we set the mean of each variable at 0, the variance of each variable at 1, and we generate a random correlation matrix using the method of canonical partial correlations suggested by Lewandowski, Kurowicka, and Joe (2010). For the first case, all variables remain continuous. We use a probit model to create binary variables for the second case, an ordered probit model to create ordinal variables for the third case, and a multinomial probit model to create unordered-categorical variables for the fourth case. Specifically, for binary variables, we turn continuous draws into probabilities using the standard normal CDF, and we generate binary values from these probabilities. For ordinal variables, we select cutpoints such that the continuous draws are divided into k equal parts, where k is the number of categories. We arrange these groupings from lowest to highest, and we assign ordered categories. For unordered-categorical variables we draw one variable for each of k categories, and set the generated category to correspond to the maximum of these values. That is, we generate a categorical value of j if the variable for category j has a higher drawn value than the values for the other $k-1$ variables. In order to preserve the independence-of-irrelevant-alternatives assumption, the variables for each category are constrained to have correlations that are only randomly different from 0 conditional on the observed and other partially observed variables.

For each variable which we intend to be partially observed, we draw values for two variables: one representing the values, one representing values of a latent missingness variable. The latent missingness variable is transformed to a standard normal probability and subtracted from a uniform$(0,1)$ random number. The cases that have the lowest 25% of these differences are selected to be missing.[10] Generating both the values and latent missingness scores together allows us to directly model MAR missingness by constraining the correlation between these two variables to be 0.

MAR can be quite difficult to impose on simulated data. Most researchers instead suppose that a few variables in the data are completely observed, and allow each variable's missingness to depend only on these fully-observed

[9] We also ran simulations with 6, 2, and no partial variables in addition to the variable of interest, but the results were not substantially different from the case with 4 additional partial variables, so these results are omitted for space.

[10] This percent is tunable, but we choose 25% as a realistic proportion of missingness for our simulations.

variables (Greenland and Finkle 1995). In this paper, we attempt to simulate a more general missingness pattern by allowing each variable's missingness to depend on binary variables indicating whether other variables are missing.

3.1 Competing MI Algorithms

We compare six approaches to missing data that fall into three categories: implementations of conditional MI, implementations of joint MI that assume a MVN distribution, and naive missing data strategies that serve as baselines for the comparisons of the other algorithms. The conditional MI algorithm is the `mi` package in R, using two different techniques to model unordered-categorical variables, the joint MVN MI algorithms are R packages `Amelia` and `Norm`, and the naive approaches are complete case analysis and draws from the empirical marginal distribution of each variable. Each of these 6 competitors is described in detail below.

We use two versions of `mi` in which continuous, binary, and ordinal variables are respectively modeled with Bayesian versions of OLS, logit, and ordered logit.[11] In one version, unordered-categorical variables are modeled with multinomial logit (MNL), as implemented in the `nnet` package (Venables and Ripley 2002). In another version, `mi` uses renormalized logit (RNL) to model unordered-categorical variables. RNL is an alternative to MNL designed to provide estimates using less computation time. Each category is modeled as a binary outcome against all other categories using a logit model, and the predicted probabilities for each category are saved. These probabilities do not in general sum to 1, so they are "renormalized," and divided by the sum of probabilities.

As discussed in section 2.1, `Amelia` and `Norm` differ primarily in how they model the variance of the imputed values. We consider both here to ensure that the results depend on the MVN approach to joint MI, and not on the method of modeling imputation variance.

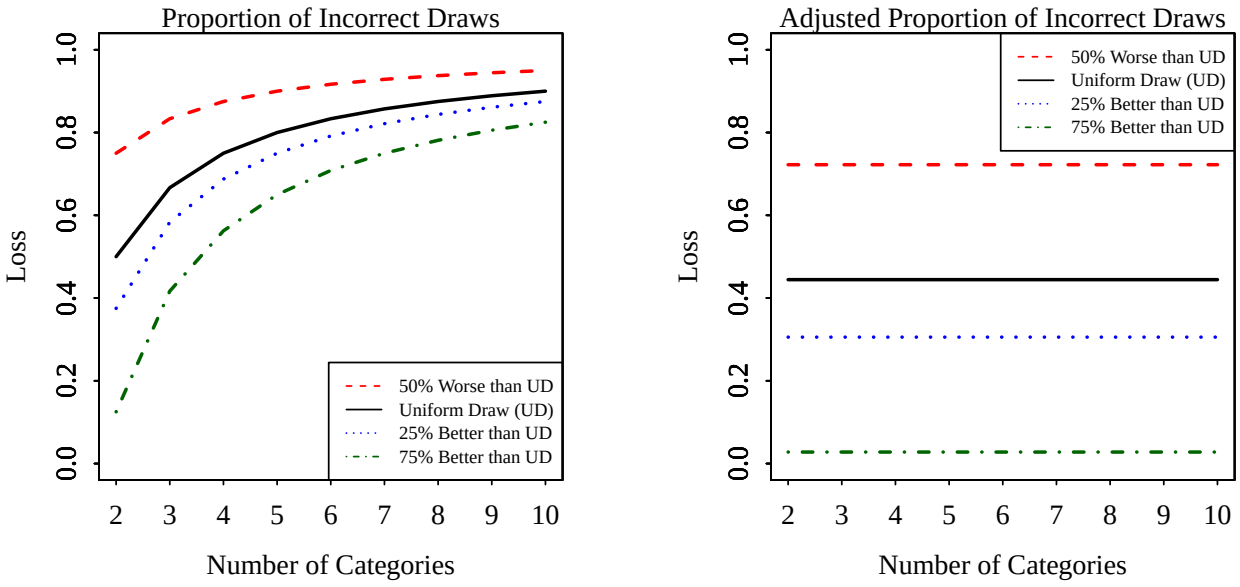
Finally, we consider CCA and draws from the empirical marginal distribution, two strategies for missing data which assume that missing values are MCAR. CCA does not impute the missing values, so it is only compared to the other competitors for its accuracy in returning the parameters of a regression after missing values have been removed. Draws from the empirical marginal distribution are sampled with replacement from the observed values of each variable. The probability that any one of the N observed cases will be selected to replace a missing value is equal to $1/N$. This replacement is independent of the values of the other variables in the data, thus this form of imputation will bias estimated associations between variables towards zero.

3.2 Evaluative Statistics

We directly compare imputed values to the true values of the missing data points, and we run a linear model on the final imputed data that we compare to the same model run on the true data. Note that the parameters of the linear model are distinct and separately estimated from the parameters of both the joint MVN imputation model

[11] These models are implemented in the `arm` package (Gelman et al 2012), and place prior distributions on the coefficients. The prior is centered around zero, which helps avoid problems of non-identified parameters due to separation (Gelman et al 2008). The prior distributions are Cauchy, and the optimum is a maximum a posteriori estimate.

Figure 1: *An Example of Adjusting the Rate of Incorrect Imputed Values to Account for the Number of Categories.*



and the conditional imputation models. We use several measures to assess to accuracy of the MI algorithms on both criteria. For every measure, lower values indicate higher quality.

Quality of Imputed Values

1. *Accuracy of imputed values.* If the variable of interest is continuous, this measure is a root mean squared error (RMSE) of the imputed values compared to the true values. If the variable of interest is binary, ordinal, or unordered-categorical, the proportion of imputed categories that match the true categories are calculated and subtracted from 1.

   For the MVN simulations, the number of categories of an ordinal variable ranges from 2 (the binary case) through 10, and the number of categories of an unordered-categorical variable ranges from 3 to 10. As the number of categories increases, the proportion of correct draws will naturally decrease. For larger numbers of categories, it is more difficult to distinguish between the high-quality and low-quality competitors.

   An example of this problem, and our solution, is illustrated in figure 1. The left-hand graph contains four generic estimators. One estimator is a random uniform draw which obtains the correct category at a rate of $1/K$, where K is the number of categories. Since we adopt a standard that lower values represent better performance, we plot $1 - (1/K)$. There are three other lines in this graph: one predicts the correct category 50% less often than the uniform draw, one predicts the correct category 25% more often than the uniform draw, and one predicts the correct category 75% more often than the uniform draw. However, although the

relative quality of the four estimators is constant, the proportions in the left-hand graph converge for higher values of K, and we are less able to draw distinctions between them. Our solution is to divide each proportion by the proportion derived by running the conditional model on true data. In the right-hand graph of figure 1, we correct the proportions by dividing them all by an estimator that predicts the correct value 80% more often than a random draw. The comparisons between the estimators are now constant across K in the right-hand graph.

2. *Accuracy of imputed choice probabilities.* Categorical imputations either return the correct category, or not. Unlike continuous imputations, it is not possible to measure the distance between imputed and true categorical values.[12] In order to get an idea of the distance between imputed and true data for categorical variables, we consider choice probabilities.

   All of the imputation algorithms estimate and use choice probabilities, although not all of the imputation algorithms intend for these probabilities to be meaningful. Conditional MI estimates the probabilities directly. We calculate probabilities from joint MVN MI, for both `Amelia` and `Norm`, using the same rules used by `Amelia` that are specifically described in section 2.1. For bootstrap draws from the marginal distribution, the probabilities are the proportion of each category in the observed data. CCA does not define probabilities or imputations for the missing values. Finally, probabilities are not defined for continuous variables, so this statistic is only calculated when we consider binary, ordinal, or unordered-categorical variables.

   We run a conditional model using the true data and save the estimated choice probability of every realized category. Then, for every imputation algorithm, we compare the corresponding probabilities to this vector of true probabilities using an RMSE measure.

Quality of a Linear Model Run On Imputed Data

Before imputation, we run a model on the true data. Each imputation algorithm outputs 5 imputed datasets,[13] and we combine estimates from the same model fit to each of these imputed datasets using Rubin's rules (Rubin 1978, 1987). We consider the following measures of model accuracy:[14]

3. *Accuracy of coefficients.* We calculate the Mahalanobis distance between the coefficient estimates from the imputed data and the coefficient estimates from the true data. We use the variance-covariance matrix of the

[12] For ordinal data it is possible to take a meaningful difference of categories, but this metric is flawed since it assumes that the ordinal categories are equally spaced.

[13] Little and Rubin (2002, ch. 10) carefully describe the statistical principles that allow a small number of drawn datasets to accurately approximate moments of the posterior distribution of the complete data. They also provide an example in which they demonstrate that "five draws of Y_{mis} can be quite adequate for generating MI inferences" (p. 211). While performing large numbers of imputations may seem appealing as an approximation to a more fully Bayesian procedure, we consider here a more general form of imputation inspired by the common situation in applied research in which several different researchers might want to each perform distinct analyses on the same study data. In this case it is helpful to be able to produce a common set of a small number of imputed datasets on which to perform a variety of different analyses.

[14] In all cases considered here, the variable of interest is the outcome in the model. For the MVN simulations, however, we calculated these statistics twice: once for the model in which the variable of interest is a predictor, and once in which the variable of interest is the outcome. When the variable of interest is a predictor, the outcome is a fully observed variable that influences whether or not the variable of interest is observed. The results in which the variable of interest is a predictor do not offer different substantive conclusions than the ones presented here, and are omitted for space.

true data coefficients for the covariance in the distance metric.

4. *Accuracy of all fitted values.* We compare the matrix of fitted values from imputed data to the matrix of fitted values from true data by taking an RMSE across all elements.[15] Since CCA does not use partial cases, it cannot provide fitted values for partial cases, so it is excluded from this comparison.

5. *Accuracy of fitted values for the complete cases, only.* This measure is equivalent to the preceding measure, except partial cases are excluded. This exclusion allows CCA to be considered relative to all of the other imputation algorithms.

Other Considerations

6. *Time to Convergence.* The time, in seconds, that passes between the start and completion of each imputation algorithm.[16]

3.3 Results

We present the MVN simulation results of the comparison of joint MVN MI, conditional MI, CCA, and bootstrap draws from the marginal distribution in this section. The results are displayed in figure 2.

Joint MVN MI and conditional MI appear to perform about equally well when the variable of interest is continuous. This result makes sense since the data are truly MVN in this case, and both conditional MI and joint MVN MI make the correct assumption about the normality of the data. However, whenever the variable of interest is categorical conditional MI outperforms joint MVN MI on every metric.

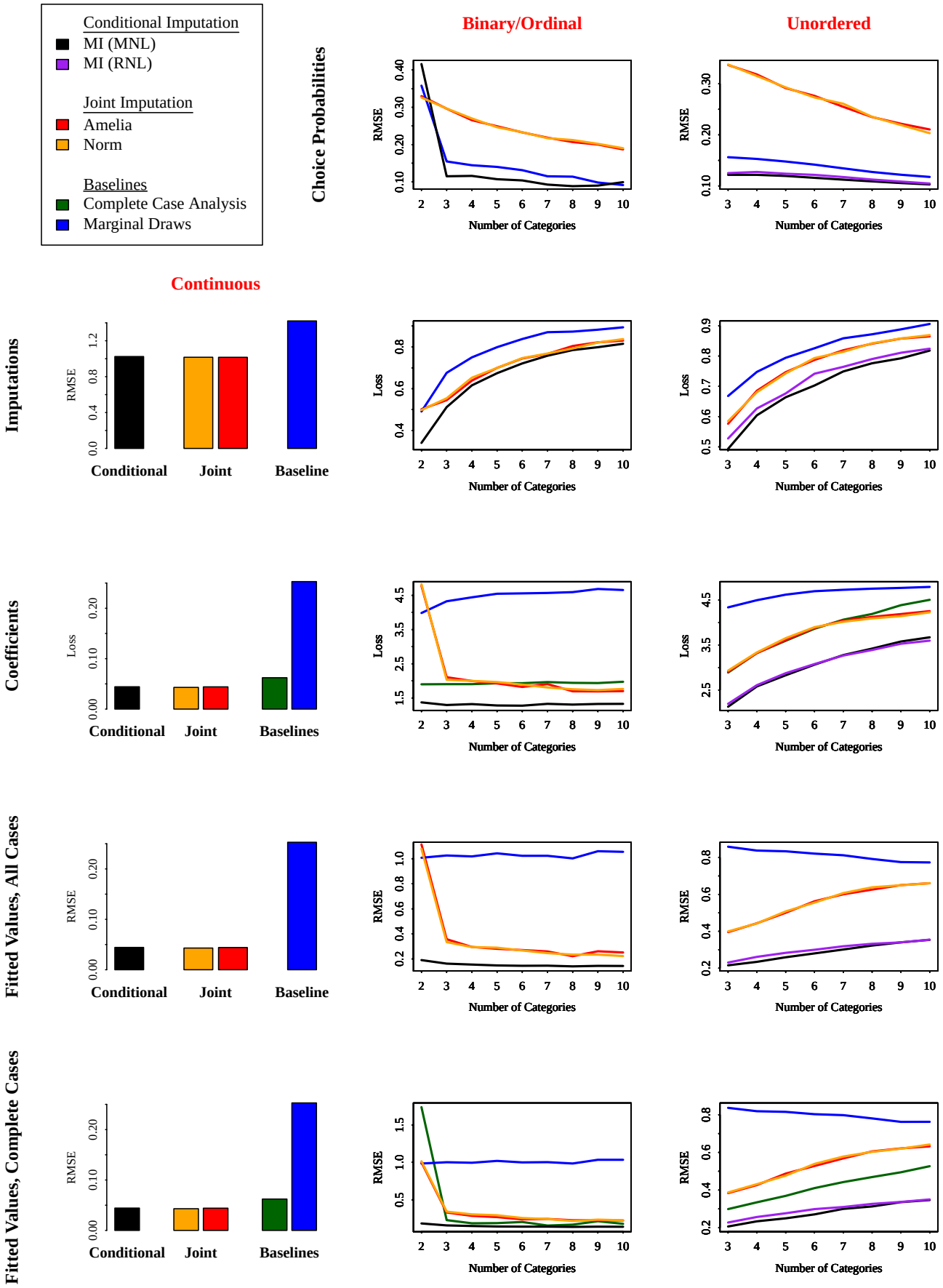
The results regarding the accuracy of choice probabilities are very similar to the results regarding the accuracy of imputed values. Conditional MI estimated binary choice probabilities poorly, but joint MVN MI performs much worse than conditional MI in estimating ordinal and unordered-categorical choice probabilities. Since these models – described in section 2.1 – provide choice probabilities for joint MVN MI that are primarily used for rounding, we expect the imputed values from joint MVN MI to be more accurate than the probabilities.

There is very little difference in the results between Amelia and Norm, so it appears that the differences between these two implementations of joint MVN MI only trivially affect their performances. Moreover, the two versions of conditional MI – the one that models unordered-categorical variables with MNL, and the one that instead uses RNL as described in section 3.1 – are very similar. Conditional MI with MNL, however, appears to consistently outperform conditional MI with RNL by a small margin. That result is in accordance with our expectations, since RNL is a less theoretical treatment of the choice probabilities. RNL, however, is much more computationally efficient

[15] For a regression and a logistic regression, the fitted values are given by the cross-product of the coefficients and the data. For ordered logit, these values are subtracted from each of the $K-1$ estimated cut-points, where K is the number of categories, and are arranged in a matrix with N rows and $K-1$ columns, where N is the number of cases. For multinomial logit, the cross-product itself forms an $N \times (K-1)$ matrix.

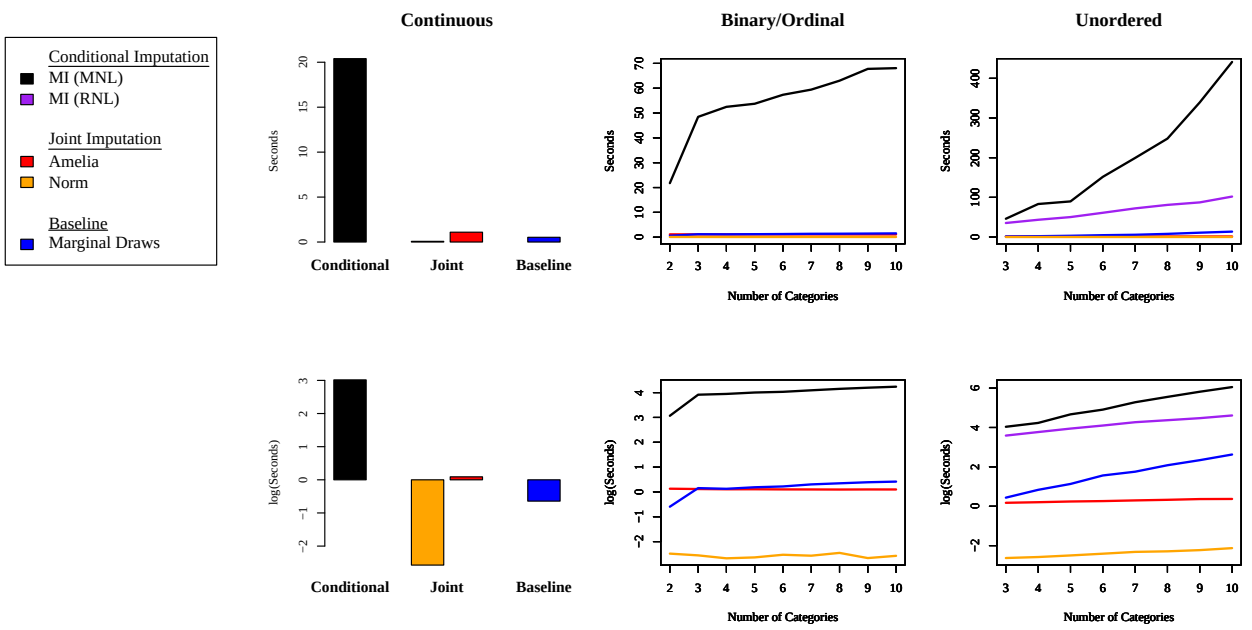
[16] Although `MI` is capable of using parallel processing, the chains are run sequentially in these simulations to reduce the amount of RAM used at any one time. The simulations are run on a series of Unix servers, each of which has 2 processors with 6 cores. The processor speed is 2.27 GHz with 54.384 GFlops. 24 GB of DDR3 memory runs at 1066 MHz.

Figure 2: *Multivariate Normal Simulation Results Using Draws from the Posterior Predictive Distribution.*



than MNL. Figure 3 shows the average time it took each imputation algorithm to converge to results and terminate

Figure 3: *Time to Convergence, Multivariate Normal Simulations.*



in the MVN simulations. Conditional MI is several orders of magnitude slower than joint MVN MI: minutes as opposed to seconds. The time required for conditional MI to run also increases with the number of categories of the categorical variable. RNL is faster than MNL, and this difference increases with the number of categories. RNL may be a viable alternative to MNL for researchers who wish to use conditional MI, but have unordered-categorical variables with so many categories that MNL is not feasible. Finally, note that in some instances CCA fails to surpass conditional MI, but still performs surprisingly well.

4 Applied Example: Simulations Based on the 2008 American National Election Study

In order to demonstrate the validity of MI on a real dataset, we use the 2008 edition of the ANES. The 2008 ANES has 1,954 variables and 2,322 cases. We impute a subset of variables that includes 2 continuous, 3 binary, 3 ordinal, and 3 unordered-categorical variables. These variables and their distributions are described in table 1. The variables are chosen with three goals in mind: first, we represent all four variable types; second, we use explanatory variables that are likely to be used in a social science study; third we formulate four plausible linear models for four interesting political outcomes. The four models considered here have the same predictors: age, sex, race, education, income, religion, and marital status. The first model regresses time on these predictors. The second model considers importance of the environment to be the outcome, and uses a logit model. The third model regresses government

Table 1: *Summary of 11 Variables in the 2008 ANES.*

Variable	Type	Missing	Distribution	Description
Age	Continuous	45	Mean=47.4 years, SD=17.4 years	Age of the respondent
Time	Continuous	225	Mean=89.3 minutes, SD=22.7 minutes	Time for the respondent to complete the survey
Importance of Environment	Binary	24	Not important: 1241; Important: 1057	Whether the respondent sees the environment as an important issue
Sex	Binary	0	Male: 999; Female: 1323	Sex of the respondent
Race	Binary	11	Non-white: 869; White: 1442	Non-white includes black, American Indian, etc.
Education	Ordinal	11	No high school: 103; Some high school: 239; High school diploma: 1476; College, plus: 493	High school diploma includes GED, some college or vocational training
Income	Ordinal	183	Low: 714; Medium: 581; High: 844	Low is below $25,000 High is above $50,000
Government Job Assistance	Ordinal	158	Extremely conservative: 363; Very conservative: 209; Conservative: 205; Moderate: 386; Liberal: 191; Very liberal: 371; Extremely liberal: 439	Responses range from “Govt should let each person get ahead on own” to “Govt should see to jobs and standard of living"
Religion	Unordered Categorical	402	Protestant: 1231; Catholic: 528; Other: 161	Other includes Jewish, atheist, Muslim, Buddhist, Hindu and other religions
Marital Status	Unordered Categorical	14	Single: 604; Married: 1013; No longer married: 691	No longer married includes divorced, separated, and widowed
Vote	Unordered Categorical	274	Obama: 1025; McCain: 514; No vote: 509	Vote choice in the 2008 Presidential election

job assistance on the predictors using a ordered logit model. The fourth model predicts respondents' votes using a multinomial logit model.

We remove the partially observed cases from the data, leaving 1,442 observations. This sample is henceforth considered to be the complete data sample. Each iteration of the simulation consists of three steps. First, missing values are generated for the complete ANES data. Then the competing MI algorithms are run on the partial data to generate imputed values. Finally the four models described above are run on the imputed data, and the results as well as the imputed values for each outcome are compared against their analogues in the complete data. Each simulation consists of 1000 iterations.

In order to make the simulation more realistic, we generate missingness that is MAR without also being MCAR. Missingness patterns are simulated using the following procedure:

1. We select one variable of each type – age, sex, income, and martial status – to remain complete. We replace

income with indicators for low income and high income, excluding medium income as a base category, and we replace marital status with indicators for single and no longer married, excluding married as the base category. We standardize these variables by subtracting their means and dividing by their standard deviations. The standardized variables form a (1442×6) matrix denoted C.

2. We take 42 independent draws from the $N(0,1)$ distribution and arrange them in a (6×7) matrix denoted β. The 6 rows of β correspond to the 6 columns of C, and the 7 columns of β correspond to the remaining variables: time, importance of the environment, race, education, government job assistance, religion, and vote. Let $\eta = C\beta$, which represents a linear combination of the columns of C for each of the 1442 cases and for each of the 7 remaining variables.

3. A matrix Z, which has the same dimensions as η, is randomly generated from a multivariate normal distribution. The columns of Z are drawn from the MVN(μ, Σ) distribution, where μ is a vector of 0s, and Σ has 1s on its diagonal, but is unconstrained for its off-diagonal elements.[17] Z is intended only to introduce correlation between the columns of η.

4. A new matrix M is constructed to be $\eta + .3Z$. The elements of M are transformed to probabilities π using the logistic distribution function.

5. For each element of π, a number $d_{i,j}$ is independently drawn from the $uniform[0,1]$ distribution, and is subtracted from $\pi_{i,j}$. In each column, the highest 10% of observations are selected to be "missing." These missingness patterns are then applied to time, importance of the environment, race, education, government job assistance, religion, and vote, respectively.

Using this method, the missingness of each partially observed variable depends both on the four complete variables and on the missingness of the other partially observed variables. We now have two versions of the data: a complete dataset, and a dataset with simulated missingness. After running MI on the partial data, we compare the imputed data to the complete data.

As with the MVN simulations, we consider the relative performances of the competing MI algorithms described in section 3.1 on the evaluative measures described in section 3.2.

4.1 Results

Figure 4 illustrates the results from the simulations that use data from the 2008 American National Election Study. Figure 4 corresponds to figure 2 which displays the results of the MVN simulations, but since these simulations do

[17] Σ is a random correlation matrix, as described by Lewandowski, Kurowicka, and Joe (2010). We mention this technique in section 3 and we describe it in greater detail in appendix 3. Unlike our use of the method in section 3, we do not include any restrictions on Σ here. Multivariate normal data with an unconstrained random correlation matrix can be generated by the `rdata.frame()` command in the `mi` package, with option `restrictions="none"`.

not increase the number of categories of ordinal and unordered-categorical variables bar charts are used instead of line charts.

We see largely the same results when we use data from the ANES as when we use MVN data. Joint MVN MI and conditional MI perform roughly equally on all metrics for the continuous variable; conditional slightly outperforms joint MVN MI for the ordinal variable; and conditional dramatically outperforms joint MVN MI for the unordered-categorical variable. Likewise, as with the MVN results, conditional MI outperforms joint MVN MI by a much greater margin for the probabilities than for the imputed values. The biggest area in which these results differ from the MVN results is for the simulations in which the outcome is binary. In these cases, joint MVN MI and conditional MI are nearly equal on the performance metrics.

Distance metrics, as evaluative statistics, do not explicitly describe how the choice of imputation method can change the conclusions we draw from a model. They demonstrate that there are real differences in the accuracy of coefficients and fitted values depending on our choice of imputation method. For true coefficients close enough to zero, a less accurate imputation method can change the sign. For true coefficients and true standard errors whose ratio is close to the cutoff for standard tests of significance, a less accurate imputation method can change the inference. Predicted probabilities are transformations of fitted values, so less accurate fitted values translate to less accurate predicted probabilities and less accurate marginal changes in probability. Whether or not these changes occur depends on the particular model and data being fit, and constitute a different question from the choice of imputation method. While conditional MI takes longer to run than joint MVN MI, the difference in computational time seems to be worth the added accuracy from conditional MI.

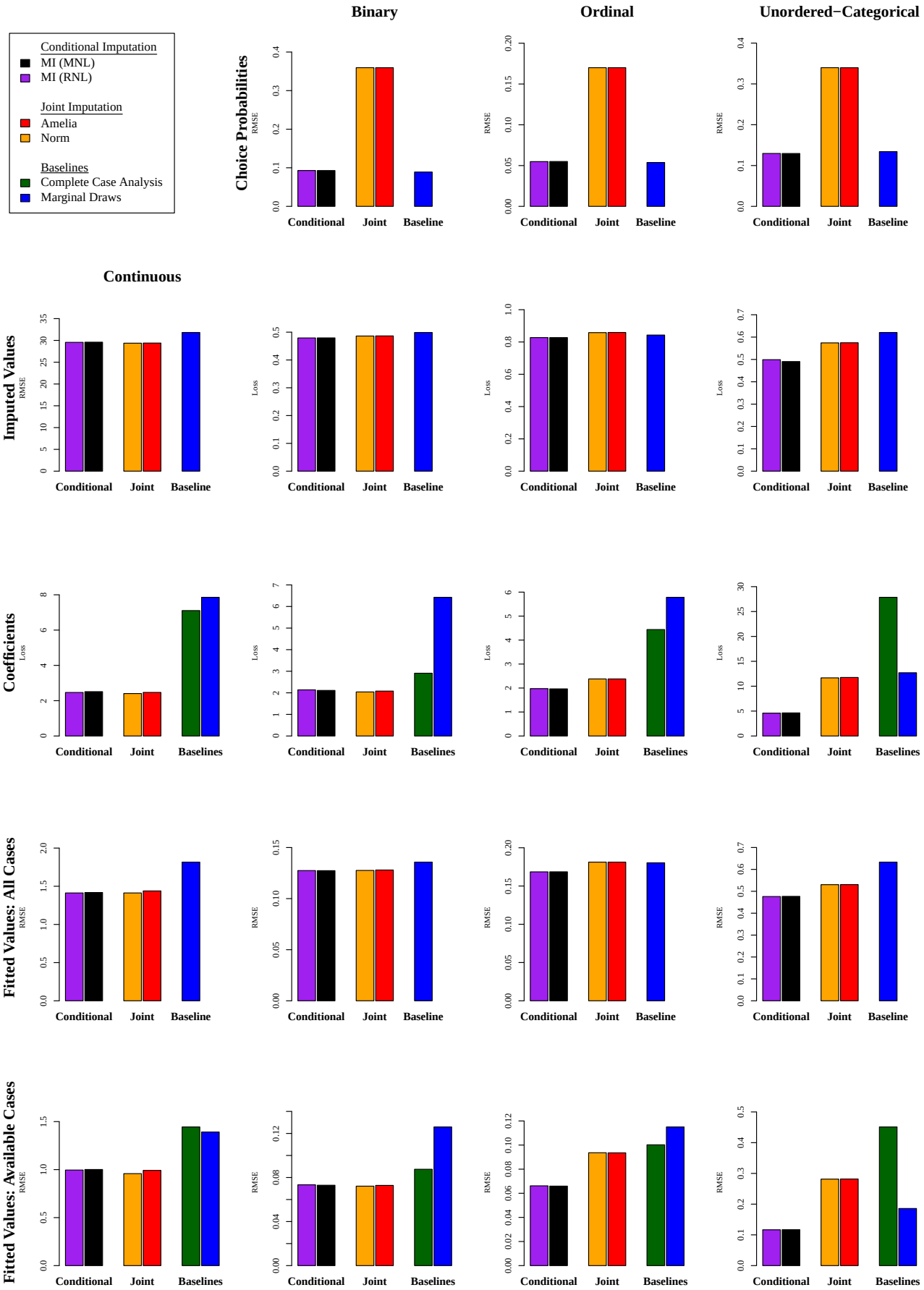
5 Discussion

In general, we see that there are two kinds of results. Either conditional MI outperforms joint MVN MI and the other competitors by a fair margin, or there is very little difference between joint MVN MI and conditional MI. In no case does joint MVN MI clearly outperform conditional MI.

We expected that two assumptions made by joint MVN MI would cause it to perform less well than conditional MI: the assumption that the data are jointly distributed MVN; and the rules used to turn categories into continuous values and continuous imputed values back into categories. We expected that joint MVN MI and conditional MI would be roughly equal only when the distributional assumption was true, and when all variables were continuous. Our expectations were largely met, although joint MVN MI exceeded our expectations in a few cases.

In the cases in which the data are MVN and every variable is continuous, joint MVN MI and conditional MI perform about equally well. We are surprised, however, that joint MVN MI and conditional MI continue to perform similarly well for continuous variables for the ANES data, which are not distributed MVN. Furthermore, joint MVN MI performed surprisingly well for the binary variable in the ANES simulations. We did not expect joint MVN

Figure 4: *Results for Simulations Based on 2008 American National Election Study Using Draws from the Posterior Predictive Distribution.*



MI to perform so well in these cases because the ANES data includes continuous, binary, ordinal, and unordered-categorical variable types, and the imputations of any one variable depend on the imputations of the other partial variables. If the imputations of an ordered or unordered-categorical variable are inaccurate, then the imputations of the continuous and binary variables should also be inaccurate. It appears, however, that joint MVN MI is resilient, and may be a viable option for imputing continuous, and perhaps even binary, variables without confirming that the data resemble a MVN distribution. This result is preliminary, and should be confirmed on simulated data that explicitly models specific kinds of non-normality.

For ordinal and unordered-categorical variables, however, conditional MI improves upon joint MVN MI for both the MVN and ANES data. For the most part, we observe an improvement in the accuracy of the imputed values themselves. But the difference is most striking for the accuracy of a generalized linear model run after imputed data are generated. Researchers using multiple imputation for the explicit purpose of keeping partially observed cases when fitting a linear model should probably consider conditional MI to be a better alternative than joint MVN MI.

6 Conclusion

It is impossible for any simulation study to consider the entire universe of possible data. We decided, therefore, to focus on two specific data structures: discretizations of MVN data and data based on the 2008 ANES. MVN data conform to the assumptions made by the joint MVN MI algorithms considered here, and should present the most favorable circumstances for joint MVN MI. The ANES is an applied example that resembles data used by many political scientists. We find that, using MVN and ANES data, joint MVN MI is faster than conditional MI, but yields imputations which are less accurate for categorical variables.

Our goal is to encourage applied researchers to choose an imputation algorithm that is appropriate for the particular characteristics of their data. While we have not conducted a universal analysis, this research should be useful to others who are considering whether to use a joint MVN or conditional implementation of MI.

We believe that two data generating processes considered here generalize to a large number of situations. However, we have not proven that the comparisons we make here will always apply for any data. In particular, we have not yet considered alternative variable types, such as truncated variables or count variables, nor have we considered alternative data structures such as multilevel data, time series cross sectional data, data with non-ignorable missingness, or high-dimensional data. Although we have no theory to suggest that joint MVN MI outperforms conditional MI in any of these settings, it may be the case that the performance of conditional MI suffers in some cases we have not examined. In future research, we intend to analyze the performance of conditional MI when some conditional distributions are explicitly misspecified.

Appendix 1: General Principles of Multiple Imputation

Rubin and Little (2002) demonstrate that complete case analysis (CCA) suffers from two problems: "loss of precision, and bias when . . . the complete-cases are not a random sample of all the cases" (p. 41). Although they suggest that these problems may be minimal when the proportion of missing data points is small and when the partial cases do not differ systematically much from the observed cases, they lay out a case that imputation is usually a better option.

In order to improve on CCA an imputation technique must have two properties. First, the technique must preserve associations between complete and partial variables and between pairs of partial variables. In other words, the imputation technique must use the observed data to make a more informed guess for each imputed value. Second, the imputed values need to be drawn from distributions rather than deterministically computed. For example, if one variable is distributed normally conditional on the other variables in the data, then replacing its missing values with the mean of this normal distribution is incorrect since it underestimates the noise in the data. Instead, the imputed values are drawn from the conditional normal distribution, which in this case serves as the posterior predictive distribution (PPD) for the partial variable (Rubin and Little 2002, p. 72). Drawing values allows several versions of the dataset to be generated. These datasets can be combined using Rubin's (1987) rules to run statistical models that contain the correct level of noise from the imputed values.

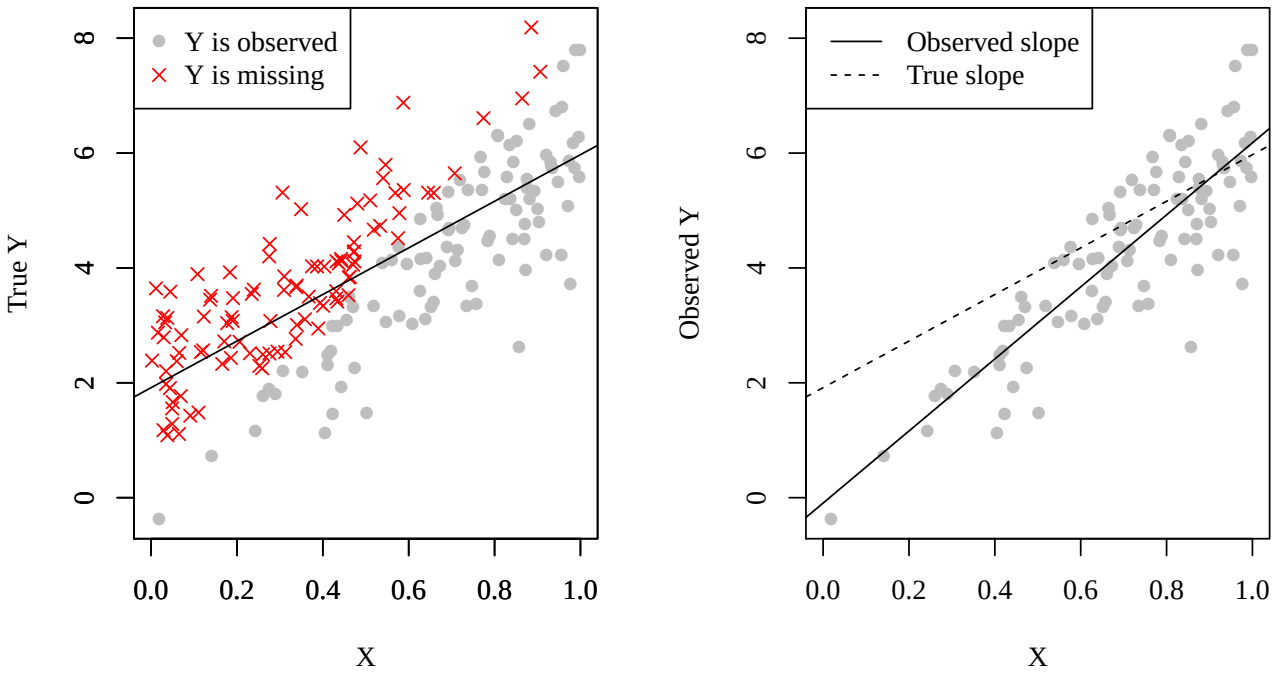
Beyond these requirements, however, Rubin and Little do not provide much guidance to researchers who must choose between contending approaches. They discuss versions of MI that assume the data jointly follow a MVN distribution and define an EM estimator. Noting that this assumption is "restrictive" (p. 224), they present some generalizations of the joint approach, including a version in which the conditional distributions are estimated through linear regression. The conditional approach itself can be generalized to model variables of different types other than continuous, and to represent a class of joint distributions larger than multivariate normal.

The Missing at Random (MAR) Assumption

Following the notation of Rubin and Little (2002, p. 12), let Y represent the complete data, where Y_{obs} denotes the observed data points and Y_{miss} denotes the missing data points. Also let M be a matrix containing the indicators of whether each data point is missing or observed, and let ϕ generally represent parameters from the joint distribution function of Y. A missing-data mechanism is missing completely at random (MCAR) if

$$f(M|Y,\phi) = f(M|\phi) \text{ for all } Y_{obs}, Y_{miss}, \phi. \tag{1}$$

Figure 5: *An Example of Inaccurate Imputation When Data are Not Missing at Random.*



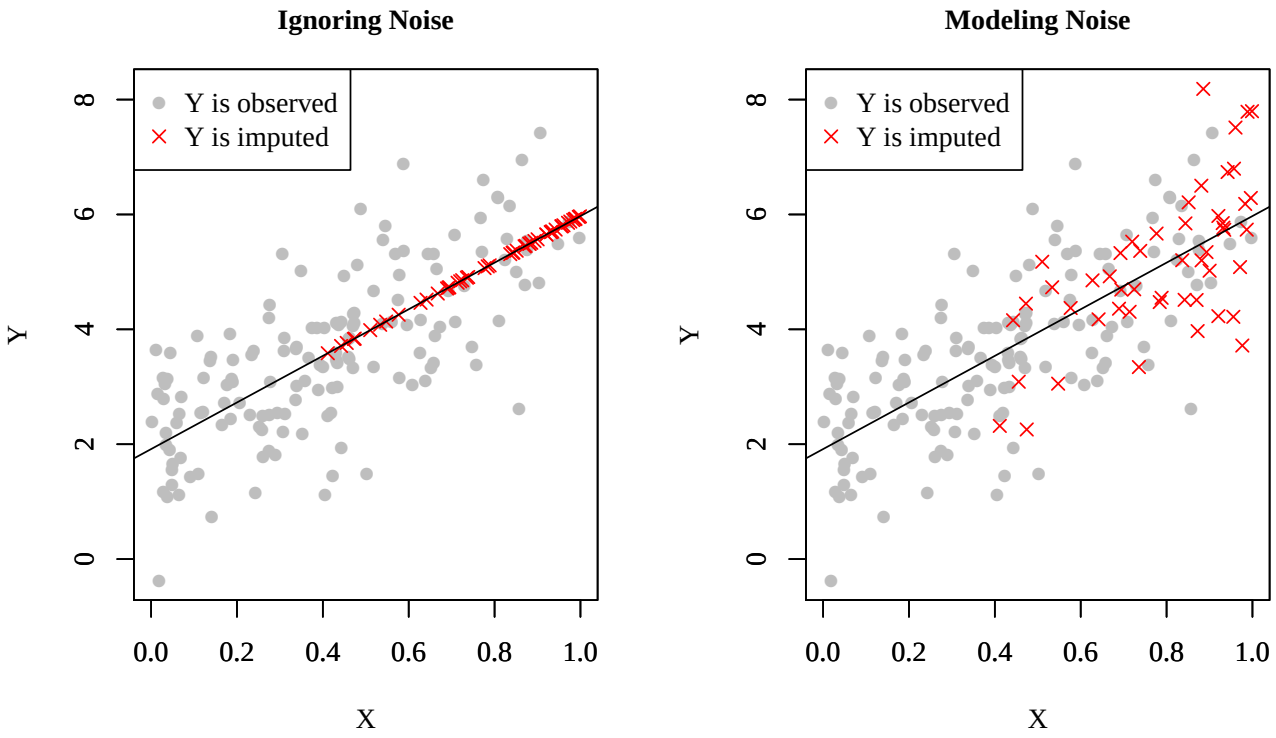
In other words, MCAR requires that whether or not a data point is missing is entirely independent of any of the real data. Less restrictive is the missing at random (MAR) condition which holds if

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \text{ for all } Y_{miss},\phi. \tag{2}$$

Here, whether a data point is missing may depend on observed data, but cannot depend upon the "true" values of missing data points for either the variable in question or for other variables. If M depends on Y_{miss}, then the data are not missing at random (NMAR), in which case MI does not provide consistent estimates. MAR and NMAR can also be thought of as graphical concepts: MAR holds when the missing data conform to the trends expressed by the observed data. Figure 5 contains a scatterplot. X is has no missing values, Y is incomplete, and whether Y is missing depends on both values of X and Y. X and Y are generated randomly, then missing points are selected, so we know the true values for the missing data points. The graph on the left plots all the true data points, and the graph on the right omits the data points that are missing on Y. Note how the best-fit line for the true data differs quite a bit from the best-fit line calculated from the complete cases.

MAR supposes that the best-fit line for the observed cases of a variable provides accurate estimates of the unobserved cases. But when the missingness is NMAR, as in figure 5, these estimates are inaccurate. CCA is inaccurate when missingness is NMAR, and no multiple imputation algorithm – regardless of the approach – can be expected to provide reliable imputations when missingness is NMAR.

Figure 6: *An Example of Imputation When Data are Missing at Random, Ignoring and Modeling Noise.*



Modeling Noise and Combining Datasets

Imputing based on observed data trends is accurate when the data are MAR, but imputations should also express the same level of noise that is present in the observed data. Figure 6 shows the difference between modeling and ignoring noise. In the graph on the left, the best-fit line for the regression of Y on X is derived, and imputed values of Y are exactly predicted from this regression. In the graph on the right, the imputed values are simulated around the predicted values using the estimated variance of the regression. Rubin and Little (2002) point out that "best prediction imputations systematically underestimate variability, [and] standard errors from the filled-in data are too small, leading to invalid inferences" (p. 64). In addition, marginal distributions and multivariate statistics such as covariances are distorted when the variance is ignored. Accurate imputation algorithms must model the noise.

There are two methods for capturing the correct noise in the imputed values. First, imputed values can be drawn from the posterior predictive distribution (PPD), by estimating the conditional distribution for each variable, simulating noise, and drawing imputations from this distribution (Rubin and Little 2002, p. 65-66). An alternative method is predictive mean matching (PMM) in which each case with a missing value is compared on some metric to every case with an observed value (Rubin 1986, Rubin and Little 2002). The imputed value is the observed data point for the closest match (Rubin and Little 2002, p. 69). Various implementations of PMM use different metrics for the comparison. One important class of MI algorithms that uses PMM is hotdeck imputation (Cranmer and Gill 2013).

Regardless of the method of modeling the noise, the imputed values within each dataset express only one possible realization of each missing data point. As a result, there is no distinction between the observed and imputed data. Imputed values allow us to use the observed data points on the same row as a missing value, but we cannot allow the imputed data to be counted as if they were observed. In order to eliminate the influence of the imputed values, several versions of the imputed data are created independently, and are combined using the rules first described by Rubin (1978, 1987). These rules adjust the results of models run post-imputation to reflect uncertainty due to variation in the imputed values. Rubin and Little (2002, p. 211-212) suggest as few as 5 imputed datasets may be sufficient to accurately describe the imputation variation.

Appendix 2: Predictive Mean Matching

In this section, we replicate our analyses using predictive mean matching (PMM) instead of drawing imputed values from the posterior predictive distribution (PPD). PPD and PMM are discussed in detail in appendix 1.

The joint MVN MI algorithms considered in this study draw imputed values using PPD by default, since imputed values are draws from the joint MVN distribution. PMM is not currently implemented in joint MVN MI software packages `Amelia` and `Norm`. For the simulations described below, we wrote original code to include PMM in these algorithms. After the MVN distribution is estimated and imputed values are drawn, we calculate the conditional mean of each observation given the values – imputed or real – of the other variables. We then match each imputed case to the observed case with the closest conditional mean. `mi` implements PMM by matching each imputed value's linear prediction from the conditional model to the closest prediction from the observed values. In order to assess probabilities using PMM, we replace the choice probabilities for missing values with the probabilities of the cases that are matched to the partial cases.

Results are presented in figures **??** and 8, which correspond to figures 2 and 4 in section 3.3. Using MVN data, joint MVN MI and conditional MI appear to perform about equally well when the variable of interest is continuous, regardless of whether PPD or PMM is used to draw imputed values. When we use PMM to draw imputed values for the ANES data, however, conditional MI outperforms joint MVN MI for every metric and variable type, including the continuous variable.

PMM is a less arbitrary method for turning continuous imputed values into categorical ones in joint MVN MI. We were surprised therefore to see that switching to PMM has very little effect on the relative accuracy of joint and conditional MI in the MVN simulations. More troubling for joint MVN MI is that in the ANES simulations, PMM causes joint MVN MI to perform less well across the board. Joint MVN MI, using PMM, still treats categorical variables as continuous, still estimates a MVN mean vector and covariance matrix, and still draws continuous imputed values. The difference is that the conditional MVN distribution for each variable is derived, given all of the other variables, and imputed categories are drawn by matching the conditional mean of each missing value to the

Figure 7: *Multivariate Normal Simulation Results Using Posterior Mean Matching.*labelresultspmm

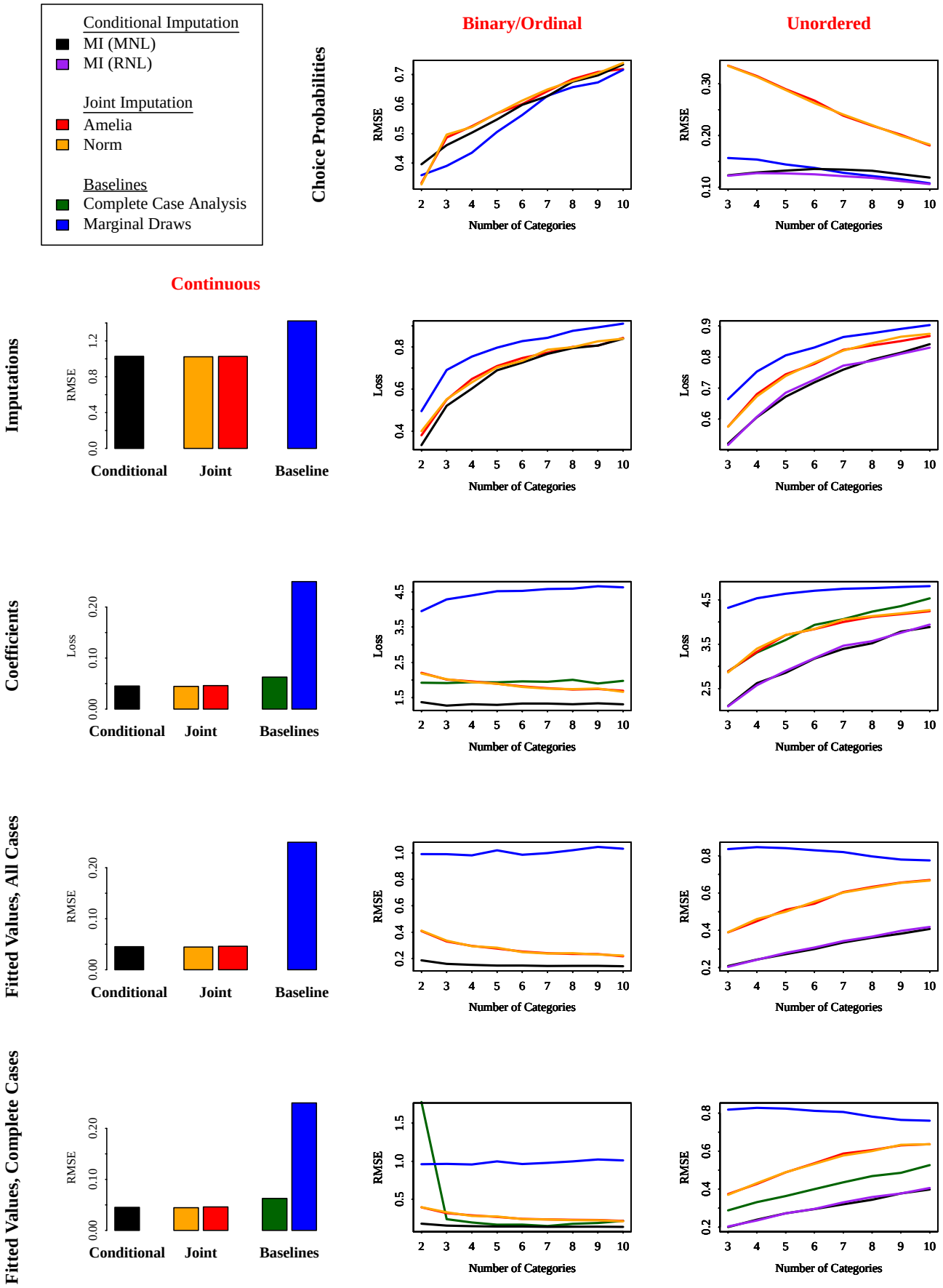
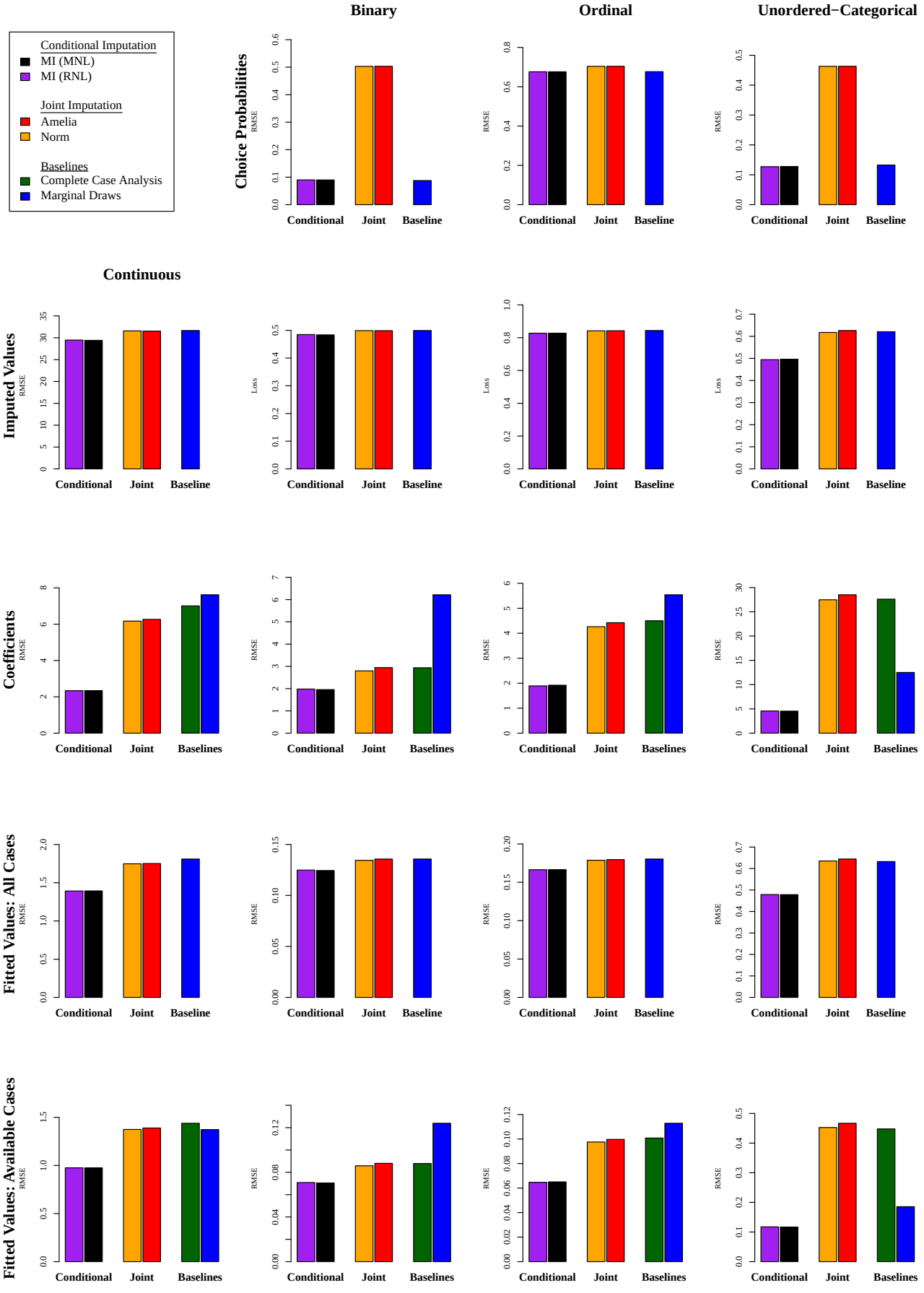


Figure 8: *Results for Simulations Based on 2008 American National Election Study Using Posterior Mean Matching.*



observed values. PMM avoids using rounding or probabilities to draw categories from continuous imputations. It does not, however, avoid treating categorical variables as continuous. Furthermore, if other variables have inaccurate imputed values, the conditional mean, and therefore matching, may be inaccurate as well. In general, these rules appear to inhibit the accuracy of joint MVN MI, and PMM – as described in section 2.1 – does not improve the performance of joint MVN MI.

Appendix 3: Generating MVN Data with MAR Missingness

One can evaluate the performance of an imputation algorithm using simulations where the data-generating process is known. However, the usual approach to Monte Carlo simulations —where complete samples are repeatedly drawn from a *fixed* population — would be suboptimal in this context for two reasons. First, it would be odd to consider the covariates to be fixed in repeated sampling when some of the observations on the covariates are missing and hence are considered random variables to be integrated out. Second, our questions of interest depend on how well an imputation algorithm performs across populations rather than across samples.

Thus, we first need to specify a distribution of population parameters. Lewandowski, Kurowicka, and Joe (2010) derive a distribution for a correlation matrix, $p(\mathbf{\Sigma}|\,\eta) \propto (\det \mathbf{\Sigma}\,(\eta))^{\eta-1}$, where $\Sigma_{ii} = 1\,\forall i$ and $\eta > 0$ is a shape parameter. By setting $\eta = 1$, the density is constant, which is to say that all correlation matrices of a given order are equally likely. It is easy to draw a correlation matrix from this distribution, and by doing so repeatedly, we could integrate over all standardized populations, knowing that the standardization does not affect any quantity of interest to us.

The next step is to distinguish between three kinds of variables in a population: fully observed variables, partially observed variables, and latent missingnesses. The vector of fully observed variables is denoted $\mathbf{X}$ and is of length n_f. The vector of partially observed variables is denoted $\mathbf{Y}$ and is of length n_p. The vector of latent missingnesses is denoted $\mathbf{Z}$ and is of length n_p because there is one latent missingness variable for each partially observed variable. All of these variables are observable in the population, but in a sample, some of the observations on each partially observed variable are missing, which depends on the sign of the corresponding latent missingness. In other words, for the ith observation in a sample, $x_{ik} = X_{ik}$, $y_{ij} = \begin{cases} Y_{ij} & \text{if } Z_{ij} < 0 \\ \text{NA} & \text{if } Z_{ij} > 0 \end{cases}$, and $z_{ij} = \text{sign}\,(Z_{ij})$.

Let $\mathbf{\Sigma} = \left[\begin{array}{c|c|c} \mathbf{\Sigma_{XX}} & \mathbf{\Sigma_{XY}} & \mathbf{\Sigma_{XZ}} \\ \hline \mathbf{\Sigma'_{XY}} & \mathbf{\Sigma_{YY}} & \mathbf{\Sigma_{YZ}} \\ \hline \mathbf{\Sigma'_{XZ}} & \mathbf{\Sigma'_{YZ}} & \mathbf{\Sigma_{ZZ}} \end{array}\right]$ be the $n \times n$ population correlation matrix partitioned according to three types of variables, where $n = n_f + 2n_p$. We could draw $\mathbf{\Sigma}$ uniformally from the distribution of $n \times n$ correlation matrices via the Lewandowski, Kurowicka, and Joe algorithm. However, such a data-generating process would be NMAR rather than MAR. To see this, assume that the population is multivariate normal and note that the conditional

covariance matrix given $\mathbf{X}$ is $\left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{YY}|\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{YZ}|\mathbf{X}} \\ \hline \boldsymbol{\Sigma}'_{\mathbf{YZ}|\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{ZZ}|\mathbf{X}} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{YY}} & \boldsymbol{\Sigma}_{\mathbf{YZ}} \\ \hline \boldsymbol{\Sigma}'_{\mathbf{YZ}} & \boldsymbol{\Sigma}_{\mathbf{ZZ}} \end{array}\right] - \left[\begin{array}{c} \boldsymbol{\Sigma}'_{\mathbf{XY}} \\ \hline \boldsymbol{\Sigma}'_{\mathbf{XZ}} \end{array}\right] \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{XY}} & \boldsymbol{\Sigma}_{\mathbf{XZ}} \end{array}\right]$. A necessary condition for MAR is that $\boldsymbol{\Sigma}_{\mathbf{YZ}|\mathbf{X}} = \mathbf{0}$, which is sufficient together with multivariate normality. Thus, in this paper where we are concerned with the behavior of imputation algorithms under MAR, we constrain the data-generating process such that $\boldsymbol{\Sigma}_{\mathbf{YZ}|\mathbf{X}} = \mathbf{0} = \boldsymbol{\Sigma}_{\mathbf{YZ}} - \boldsymbol{\Sigma}'_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XZ}}$.

If the data were not multivariate normal, then it would not necessarily be correct to describe the data-generating process as MAR even if $\boldsymbol{\Sigma}_{\mathbf{YZ}|\mathbf{X}} = \mathbf{0}$. However, most statistical analyses hinge on the first two moments of the data, in which case we would not anticipate the biases to be substantial if $\mathbf{Y}$ and $\mathbf{Z}$ were orthogonal given $\mathbf{X}$ but had some dependence in the higher moments.

Our population correlation matrices are not quite uniform conditional on $\boldsymbol{\Sigma}_{\mathbf{YZ}|\mathbf{X}} = \mathbf{0}$, but they are more general than the usual data-generating processes in the missing data literature. Most of the literature implicitly generates data where $\boldsymbol{\Sigma}_{\mathbf{ZZ}|\mathbf{X}}$ is not only fixed but is constrained to be diagonal, which is not a requirement for MAR. Indeed, one of the salient features of most real samples is that missingness is clustered across variables such that if an observation is missing on one variable, it tends to be missing on another variable. In our terminology, $\boldsymbol{\Sigma}_{\mathbf{ZZ}}$ would have large, positive, off-diagonal elements, and this dependence among the latent missingnesses would persist even after conditioning on $\mathbf{X}$.

We have conducted simulations with and without the restriction that $\boldsymbol{\Sigma}_{\mathbf{YZ}|\mathbf{X}} = \mathbf{0}$ but report the results where this unrealistic restriction is imposed. The joint MVN imputation algorithms specify a multivariate normal distribution over $\mathbf{X}$ and $\mathbf{Y}$ and ignore $\mathbf{Z}$. A conditional imputation algorithm can condition on sign $(\mathbf{z}_{-j})$ when imputing a missing y_{ij}. If the population were generated such that $\boldsymbol{\Sigma}_{\mathbf{YZ}|\mathbf{X}} \neq \mathbf{0}$, we expect conditional imputation algorithms to perform better than joint MVN imputation algorithms, although in practice the difference seems to be small at most. Thus, we report the results where the population is generated such that $\boldsymbol{\Sigma}_{\mathbf{YZ}|\mathbf{X}} = \mathbf{0}$, which is somewhat disadvantageous to a conditional imputation algorithm because conditioning on sign $(\mathbf{z}_{-j})$ as well as $\mathbf{x}$ requires estimating additional coefficients that are zero in the population.

The literature also usually generates a population where $\boldsymbol{\Sigma}_{\mathbf{YY}|\mathbf{X}}$ is not only fixed but is constrained to be diagonal, which is also not a requirement for MAR. We do not impose this restriction on our random populations, although $\boldsymbol{\Sigma}_{\mathbf{YY}|\mathbf{X}}$ is diagonal in expectation. Although allowing $\boldsymbol{\Sigma}_{\mathbf{YY}|\mathbf{X}}$ to be non-diagonal is more realistic, we do not expect it to have much affect on the simulations because neither the joint MVN nor the conditional imputation algorithms assume anything about $\boldsymbol{\Sigma}_{\mathbf{YY}|\mathbf{X}}$. If $\boldsymbol{\Sigma}_{\mathbf{YY}|\mathbf{X}}$ were sparse and the sparsity pattern were somehow known, then perhaps a conditional imputation algorithm could gain an advantage over a joint MVN imputation algorithm by imposing the corresponding exclusion restrictions, but we do not accept the premise that the sparsity pattern would be known in general. The literature often, although not always, generates a population where $n_p = 1$, which is both unrealistic and unlikely to provide a good basis for comparing imputation algorithms. In our simulations, we manipulate n_p.

Thus, it is possible to draw $\mathbf{\Sigma}$ such that $\mathbf{\Sigma}_{\mathbf{YZ}|\mathbf{X}} = \mathbf{0}$ and draw a "true sample" of size N from a multivariate normal distribution with mean vector zero and correlation matrix $\mathbf{\Sigma}$. We then construct an "observed sample" from the true sample by changing y_{ij} to NA iff $z_{ij} > 0$ and then discarding all the latent missingnesses. Finally, we apply imputation algorithms to the observed sample and calculate various loss functions. It is quite possible to introduce skewness or kurtosis, but we do not do so in this paper.

When some of the variables are discrete, the process of constructing an observed sample is more involved. If a fully observed or partially observed variable is ordinal, we take the corresponding marginal from the multivariate normal distribution and discretize it. The cutpoints are chosen such that the probability of falling in each category is equal. This restriction on the cutpoints is not particularly realistic but avoids the potential situation the sample is small, the number of categories is large, and some category is empty in the observed sample. We manipulate the number of categories in our simulations. A binary variable is just a special case of an ordinal variable with two categories, but in that case we simply fix the cutpoint at zero.

Constructing a nominal variable is more complicated than an ordinal variable. An observation on a nominal variable with K categories can be generated from K latent variables such that the nominal value takes the kth level if the kth latent variable is larger than the other $K-1$ latent variables. Thus, we generate these K latent variables in the population and make them conditionally independent given the fully observed variables and the previous partially observed variables. In the true sample, only the nominal values are observed.

References

The American National Election Studies (ANES; www.electionstudies.org). The ANES 2008 Time Series Study [dataset]. Stanford University and the University of Michigan [producers].

Bernaards, Coen A., Thomas R. Belin, and Joseph L. Schafer. 2007. "Robustness of a Multivariate Normal Approximation for Imputation of Incomplete Binary Data." *Statistics in Medicine*. 26(6): 1368-1382.

Cranmer, Skyler J. and Jeff Gill. Forthcoming. "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data." *British Journal of Political Science*.

Cribari-Neto, Francisco and Achim Zeileis. 2010. "Beta Regression in R." *Journal of Statistical Software*. 34(2): 1-24.

Demirtas, Hakan. 2010. "A Distance-Based Rounding Strategy for Post-Imputation Ordinal Data." *Journal of*

Applied Statistics. 37(3): 489-500.

Dempster, Arthur P., Nan Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)*. 39(1): 1-38.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *Annals of Applied Statistics*. 2(4): 1360-1383.

Gelman, Andrew, Yu-Sung Su, Masanao Yajima, Jennifer Hill, Maria Grazia Pittau, Jouni Kerman and Tian Zheng. 2012. "arm: Data Analysis Using Regression and Multilevel/Hierarchical Models." R package version 1.5-05. <http://CRAN.R-project.org/package=arm>

Goodrich, Ben, Jonathan Kropko, Andrew Gelman, and Jennifer Hill. 2012. "mi: Iterative Multiple Imputation from Conditional Distributions." R package version 2.15.1.

Greenland, Sander and William D. Finkle. 1995. "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses." *American Journal of Epidemiology*. 142(12): 1255-1264.

Honaker, James and Gary King. 2010. "What to do About Missing Values in Time Series Cross-Section Data." *American Journal of Political Science*. 54(2):561-581.

Honacker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software*. 45(7): 1-47.

—. 2012. "Amelia II: A Program for Missing Data." Software documentation, version 1.6.2. <http://r.iq.harvard.edu/docs/amelia/amelia.pdf>.

Horton, Nicholas J., Stuart R. Lipsitz, and Michael Parzen. 2003. "A Potential for Bias When Rounding in Multiple Imputation." *The American Statistician*. 57(4): 229-232.

Lee, Katherine J. and John B. Carlin. 2010. "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation." *American Journal of Epidemiology*. 171(5): 624-632.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2010. "Generating Random Correlation Matrices Based

on Vines and Extended Onion Method." *Journal of Multivariate Analysis*. 100(9): 1989-2001.

Li, Fan, Yaming Yu, and Donald B. Rubin. 2012. "Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines." Working paper. <ftp.stat.duke.edu/WorkingPapers/11-24.pdf>. Accessed 7 December 2012.

Royston, Patrick. 2005. "Multiple Imputation of Missing Values: Update." *Stata Journal*. 5(2): 188-201.

—. 2007. "Multiple Imputation of Missing Values: Further Update of Ice, with an Emphasis on Interval Censoring." *Stata Journal*. 7(4): 445-474.

—. 2009. "Multiple Imputation of Missing Values: Further Update of Ice, with an Emphasis on Categorical Variables." *Stata Journal*. 9(3): 466-477.

Rubin, Donald B. 1978. "Multiple Imputations in Sample Surveys." Proceedings of the Survey Research Methods Section of the American Statistical Association.

—. "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations." *Journal of Business and Economic Statistics.* 4(1): 87-94.

—. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Rubin, Donald B. and Roderick J. A. Little. 2002. *Statistical Analysis with Missing Data*. Second ed. New York: John Wiley and Sons.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, Joseph L. and Maren K. Olsen. 1998. "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective." *Multivariate Behavioral Research*. 33(4): 545-571.

StataCorp. 2013. *Stata 13 Base Reference Manual*. College Station, TX: Stata Press.

Su, Yu-Sung, Andrew Gelman, Jennifer Hill, and Masanao Yajima. 2011. "Multiple Imputation with Diagnostics (mi) in R: Opening Windows Into the Black Box." *Journal of Statistical Software*. 45(2).

Therneau, Terry. 2012. "survival: A Package for Survival Analysis in S." R package version 2.36-14.

van Buuren, Stef. 2007. "Multiple Imputation of Discrete and Continuous Data By Fully Conditional Specification." *Statistical Methods in Medical Research.* 16(3): 219-242.

—. 2012. *Flexible Imputation of Missing Data.* Boca Raton, FL: Chapman & Hall/CRC.

van Buuren, Stef, Hendriek C. Boshuizen, and D. L. Knook. 1999. "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine.* 18(6): 681-694.

van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software.* 45(3).

Venables, William N. and Brian D. Ripley. 2002. *Modern Applied Statistics with S.* Fourth Edition. New York.

Yu, L-M, Andrea Burton, and Oliver Rivero-Arias. 2007. "Evaluation of Software for Multiple Imputation of Semi-Continuous Data." *Statistical Methods in Medical Research.* 16(3): 243-258.

Yuan, Yang C. 2013. "Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0)." SAS Software Technical Papers.