Principled Imputation Made Simple: Multiple Imputation Using Gaussian Copulas[*]

Florian M. Hollenbach[†]
Department of Political Science, Texas A&M University
and
Iavor Bojinov
Department of Statistics, Harvard University
and
Shahryar Minhas
Department of Political Science, Michigan State University
and
Nils W. Metternich
Department of Political Science, University College London
and
Michael D. Ward
Department of Political Science, Duke University
and
Alexander Volfovsky
Department of Statistical Science, Duke University

May 17, 2017

[*]Florian M. Hollenbach is an Assistant Professor, Department of Political Science, Texas A&M University, College Station, TX 77843-4348 (email: fhollenbach@tamu.edu); Iavor Bojinov is a PhD Student, Department of Statistics, Harvard University, Cambridge, MA 02138 (email: bojinov@fas.harvard.edu); Nils W. Metternich is a Senior Lecturer, Department of Political Science, University College London, London, UK WC1H 9QU (email: n.metternich@ucl.ac.uk); Shahryar Minhas is an Assistant Professor, Department of Political Science, Michigan State University, East Lansing, MI, 48824 (email: s7.minhas@gmail.com); Michael D. Ward is a Professor, Department of Political Science, Duke University, Durham, NC 27708 (email: michael.d.ward@duke.edu); and Alexander Volfovsky is an Assistant Professor, Department of Statistical Sciences, Duke University, Durham, NC 27708 (email: av136@stat.duke.edu). This project was partially supported by the the Office of Naval Research (holding grants to the Lockheed Martin Corporation, Contract N00014-12- C-0066). Nils W. Metternich acknowledges support from the Economic and Social Research Council (ES/L011506/1). The work was completed while Alexander Volfovsky was supported by a NSF MSPRF under DMS-1402235. For helpful insights we thank Philippe Loustaunau, among the first of our colleagues to encourage this effort. Stephen Shellman was a strong critic who deserves our thanks too: his criticisms helped us to improve our approach. John Ahlquist, Matt Blackwell, Andreas Beger, Cassy Dorff, Gary King, and Jacob Montgomery provided helpful comments on previous versions of this paper.

[†]Corresponding author

Abstract

Missing observations are pervasive throughout observational research, especially in the social sciences. Despite multiple approaches to dealing adequately with missing data, many scholars still rely on list-wise deletion. In this article, we present a simple to use approach to multiple imputation. We show that using Gaussian copulas for multiple imputation allows scholars to attain estimation results that have good coverage and small bias. Using simulated as well as observational data from published social science research we compare imputation via Gaussian copulas with two other widely used imputation methods: MICE and Amelia II. The three approaches perform relatively similarly. Importantly, however, imputation via the Gaussian copula is simple and does not require the researcher to undertake any transformation of the data or specification of distributional assumptions for individual variables but returns a valid posterior density of the imputed data.

Keywords: missing data, Bayesian statistics, categorical data

1 Introduction

Missing data problems are ubiquitous in observational data and common among social science applications. Statistical inference that does not adequately account for the missing data can lead to biased results and inflated (or deflated) variance estimates (Rubin, 1976, King et al., 2001, White and Carlin, 2010, Molenberghs et al., 2014). Despite these well established results many applied scientists still choose to ignore this problem. Principled approaches to missing data have existed for over three decades, and since their formalization in Rubin (1976) the number of readily available statistical software to apply them have rapidly grown (King et al., 2001, Honaker and King, 2010, Van Buuren and Groothuis-Oudshoorn, 2011, Kropko et al., 2014) (See also the special issue on the *State of Multiple Imputation Software* in the *Journal of Statistical Software* in 2011 (Yucel, 2011)). Most often these software use multiple imputation (MI) to fill in the missing data and create m completed data sets, which are analyzed independently and combined to obtain a final estimate that accounts for the uncertainty due to the missing data (Rubin, 1996, 2004).

Even though imputation packages exist in almost any statistical software platform they are still under utilized. Van Buuren (2012) surveys multiple imputation approaches and his website (`http://www.stefvanbuuren.nl/mi/index.html`) suggests that the number of publications with “multiple imputation” in the title is growing exponentially from 1990-2011, but is still only a few dozen each year. We believe this is due to two reasons. First, imputation techniques are not well understood and can produce results that seem counter-intuitive. Graham (2009) highlights a number of myths that discourage people from using imputation methods in practice. For example, many people still believe that one cannot impute the dependent variable or are unsure what to do if the MAR assumption does not hold. Second, imputation techniques are burdensome and require three stages: data imputation, model estimation, and combining estimation results. Particularly, the specification and estimation of the imputation model can be time consuming and computationally expensive. This poses an enormous problem with the increasing prevalence of “big data” as the issues of model specification and computational cost become more significant as sample size and the number of variables increase.

For example, one of the most commonly used multiple imputation techniques, multiple imputation via chained equations (MICE), requires users to specify the “correct” imputation model (i.e. probability distribution and predictors) for each variable that is to be imputed (Van Buuren and Groothuis-Oudshoorn, 2011). Only if *each* specification is correct, is MICE guaranteed to converge to a valid joint distribution. Similarly, when using Amelia II (Honaker et al., 2013), another widely used multiple imputation method, users should specify the correct variable types and transformation in their data set and a number of tuning parameters.

To tackle these problems our paper presents a semi-parametric Gaussian copula approach to missing data imputation. Copulas were first used for imputation by Käärik and Käärik (2009), who introduce the use of Gaussian copulas to model missing values based on the observed data. Recently, Di Lascio et al. (2015) have shown how copulas from several different distributions can be used for imputations and compare their approach to nearest neighbor and regression imputation. Within sociology, Vuolo (2015) presents the use of copulas for modeling of join distributions. Nevertheless, the potential use of copulas for multiple imputation applications has not been thoroughly discussed, nor has it been widely adopted, especially within the social sciences. The method we present here was developed by Hoff (2007) and implemented in an **R** package (Hoff, 2010). Based on the use of the rank-likelihood, the method presented here has additional advantages over those previous ones as it allows for the imputation of binary and ordinal variables and does not require the specification of marginal or conditional distributions.

In this paper, we provide an overview and discussion of common imputation methods and show that the semi-parametric Gaussian copula model is easily usable for multiple imputation. For example, it does not require any prior specification by the user or preparation of the data. Moreover, provided the MCMC chains converge, the output from the copula model represents a valid posterior density. In addition, we provide guidance on how to evaluate different multiple imputation techniques. The accuracy of imputations is often judged based on mean squared error between the "true" and imputed values. As we elaborate below, this can be misleading. Instead, multiple imputation ought to be evaluated on bias, coverage rate, and interval length for the regression coefficients of interest (Rubin, 1996).

The remainder of this article is organized as follows. Section 2, discusses some common approaches to missing data. Section 3, thoroughly explains the Gaussian copula approach and how it can be utilized to impute missing data. Section 4, presents a large simulation study that assesses the properties of our proposed approach and compares it to some of the most commonly used procedures. Section 5, applies the methods from Section 4 on a real data example. Section 6, gives our concluding remarks as well as some guidance for practitioners.

2 Common Approaches to Multiple Imputation

The standard techniques employed to deal with missing data first require an assumption regarding the observed missing data pattern, Rubin (1976, 1987a,b), Little and Rubin (2002) introduced three of these. The missing data are missing completely at random (MCAR) when the probability of the observed missing data pattern is unchanged regardless of what values both the observed and missing data take. The missing data are missing at random (MAR) when the probability of observing the missing data pattern is unchanged no matter what values the missing data take. Finally, the missing data are missing not at random (MNAR) when the probability of observing the missing data pattern changes for some values of the missing data.

These definitions are important both from a theoretical and a practical point of view. Most basic methods, such as list-wise deletion and mean imputation, require the MCAR assumption (Graham, 2009). To achieve valid inference under the Bayesian and likelihood paradigms, whilst ignoring the missing data mechanism, we require the stronger MAR assumption. Horton and Kleinman (2007) provide a good overview of some valid procedures including using the expectation maximization (EM) algorithm and chained equations (Van Buuren, 2012) to obtain multiple imputations (MI) (Rubin, 1987b). Unfortunately, as Horton and Kleinman (2007) conclude, there is little evidence to suggest that these techniques are widely employed by analysts in medicine and the social sciences.

Multiple imputation (MI) refers to any method that replaces the set of missing values with multiple plausible values, thus obtaining m completed data sets (Rubin, 1996). It has been applied in many fields, but the most well-known application is in survey sampling. In this setting, we use the responses of those who answered a particular set of questions to provide an estimate of the response of those who did not answer a particular question. One approach to generating MI is to use a joint multivariate model for the observed and missing data, and then draw from the posterior distribution of the missing data given the observed data. As the name indicates, multiple imputation relies on the strategy of generating multiple imputed observations for each missing data point, resulting in the creation of several complete data sets. Rubin (1987a) originally suggested creating five imputations, but more recently authors recommended using closer to twenty imputations (Van Buuren, 2012). These data sets are then separately analyzed using the standard full data techniques, the resulting quantities of interest from each data set are then combined to obtain an overall estimate as well as its associated variance.

MI with EM Approaches that use iterative expectation maximization (EM) to create complete data sets were originally developed by Dempster et al. (1977), but more recently advanced by Schafer (1997) and Honaker and King (2010). These methods model the joint distribution as a multivariate normal distribution. Honaker and King (2010) provided an implementation of this method in R, known as Amelia, that combines the EM approach with bootstrapping to derive solutions more quickly and runs in parallel. In large data sets with significant amounts of missing data the package can be computationally intensive, a trait that is a characteristic of EM algorithm as the rate of convergence is proportional to the amount of missing information in the model.

Conditional Approaches to Multiple Imputation An alternative technique is to model each variable's imputation via its conditional distribution based on all other variables in the model. One such approach is Multiple Imputation via Chained Equations (MICE) (Van Buuren, 2012). Imputations for fully conditional specification (FCS) methods, such as MICE, in general work by first starting with an initial guess of Y^{mis} (*e.g.* the mean). This "imputed" variable is used as dependent variable (with missing values restored) in a regression on all other variables. The regression estimates are used to impute the missing observations, based on the patterns among the observed values (including those imputed) among the variables included in the modeling exercise. The estimates and parameters are then updated by cycling through the variables and imputing each one given the most current estimate of the parameter and other variables. The procedure stops when the chain has converged. A similar technique is used in the *MI* package in $\mathcal{R}$ (Goodrich et al., 2012). Each variable is imputed based on an "appropriate generalized linear model for each variable's conditional distribution" (Kropko et al., 2014, 501). This is done for all variables and iterated until the model converges.

> One of the main drawbacks of the FCS is that they do not necessarily define a valid joint distribution and therefore can lead to pathologies in the convergence of the algorithms (Li et al., 2012). Liu et al. (2013) showed that for valid semicompatible models (*i.e.* models which are compatible when some of the parameters in the conditional distributions are set to zero, and the joint model obtained from the compatible conditionals contains the true joint probability distribution) the combined imputation estimator is consistent. Further, Zhu and Raghunathan (2015) extend these results to more incompatible models at the expense of the type of missingness patterns allowed (restricting the theoretical results to missingness patterns where each individual is missing at most one variable). Beyond these theoretical developments, one of the advantages of conditional model specification is that it allows each variable to be modeled based on its specific distribution, which is specified by the researcher. However, this can be "labor-intensive and challenging with even a moderate number of variables" (Murray, 2013, 41). Moreover, coefficients estimates in the conditional models can suffer significantly when the number of missing observations is large, especially for categorical variables (Murray, 2013).

In the next section we describe an alternative approach, first introduced by Hoff (2007), where imputation is done via MCMC methods utilizing Gaussian copulas. For the remainder of this article we the focus on a comparison of the copula based imputation method to the MICE algorithm for FCS methods and Amelia for EM methods, based on their corresponding R implementations (Van Buuren and Groothuis-Oudshoorn, 1999, Van Buuren and Groothuis-Oudshoorn, 2011, Blackwell et al., 2015).

3 A copula approach to missing data imputation

One of the key issues with conditional approaches to imputation, such as MICE, is that they do not necessarily specify a valid joint distribution. In turn this cannot guarantee the proper behavior of confidence intervals and overall inference. A natural approach to overcoming possibly incompatible conditional specification is by specifying the joint distribution directly. Since this problem becomes increasingly complicated as the number of covariates in the model increase, it is valuable to decouple the specification of the marginal distribution of each covariate from the function that describes the joint behavior of all covariates. One such function is called a Copula and Sklar's (1959) theorem guarantees that every joint distribution can be decomposed in this way:

Theorem 3.1 (Sklar's Theorem)**.** *Let F be a n-dimensional joint distribution function with marginals $F_1, \ldots, F_n$. Then there exists a copula C with uniform marginals such that*

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n))$$

Much work has been done studying the class of Gaussian copulas where the multivariate dependence is defined by C via the multivariate normal distribution with a correlation matrix R (Klaassen et al., 1997, Pitt et al., 2006, Chen et al., 2006, Hoff, 2007). Of particular interest is the setting where no parametric form is specified for the marginal distributions $F_1, \ldots, F_n$, making this a semiparametric approach. In this flexible setting, estimation procedures are still equipped with theoretical guarantees for the parameters of the copula model (Murray et al., 2013, Hoff et al., 2014). Since these parameters determine the dependence structure, and so direct the imputation of the missing data, these theoretical results are extremely appealing. The estimation approach we explore below was developed by Hoff (2007) by extending the ideas of the rank likelihood Pettitt (1982) to the copula setting.

Using the notation of Hoff (2007), data that is generated by a multivariate Gaussian copula can be written as $y_{ij} = F_j^{-1}(\Phi(z_{ij}))$ where $z_1, \ldots, z_n \overset{iid}{\sim} \mathcal{N}(0, R)$ with R a correlation matrix and F_j the univariate cumulative distribution function (CDF) of variable j. The rank likelihood (Pettitt, 1982), a type of marginal likelihood that bases inference on the ranks of data rather than the full data, leverages the ordering of the observed values $y_{1j}, \ldots, y_{nj}$ of each variable to make inference about the parameter R without estimating the CDFs $F_1, \ldots, F_p$.

A Bayesian approach to estimating R specifies an inverse Wishart prior for a covariance matrix V such that R is its correlation matrix and a normal prior for the latent z_{ij}. Updates are performed via a Gibbs sampler as full conditional distributions can be derived by conditioning on the ranks of the data alone. Details of the algorithm for estimation are available in Hoff (2007) and are implemented in an **R** package (Hoff, 2010).

When values of y_{ij} are missing at random, imputation can be performed first on the latent z_{ij} scale and then transformed to the observed scale using the empirical CDFs. As this is a Bayesian procedure we produce a full posterior for the missing data. To make our approach comparable to the standard conditional approaches we only employ a few samples from this posterior and use those as multiply-imputed datasets. However, it is natural to consider posterior predictive distributions of parameters of interest or other posterior summaries on a case-by-case basis. For example, the conditional independence graphs of Hoff (2007) succinctly summarize the relationships among many variables.

4 Comparing Amelia II, Copula, and MICE

In this section we compare copula based imputation (Hoff, 2007) with some of the commonly used `Amelia II` and `MICE` packages. We evaluate each technique based on a simulation study as well as a real world data application from the social sciences, discussed in the next section.

4.1 Evaluating Imputations

Multiple imputation procedures are specifically designed to yield valid statistical inference (meaning, asymptotically unbiased with correct standard errors and coverage) for population quantities of interest. The name itself suggests that each missing data point will be imputed multiple times, taking on possibly different values in each completed data set. As such, simulation based evaluation of the efficacy of a multiple imputation procedure based on deviation (squared, absolute, or otherwise) from the obscured data is likely to be misleading (Rubin, 1996, 2004). For example, consider the standard problem of minimizing mean squared error – the best estimator in this case is the mean which is agnostic to other information and will be lead to an underestimate of the standard errors. Since correct estimation of the standard errors is critical for obtaining valid statistical inference any analysis of the MI procedure must focus on studying its frequentist properties. Properties such as empirical coverage, average bias and average interval length of the estimate of the scientific estimand over repeat samples will be of cardinal interest.

Assessing MI: To assess the validity of a MI procedure through simulation we use the following approach:

1. Define a full data quantity of interest, θ. For example, this could be a set of regression coefficients.

2. Generate a complete data set and apply a pre-specified missing data mechanism to remove some observations.

3. Use the MI procedure to create m completed data sets with the missing values replaced by imputed values.

4. Use each of the m data sets to obtain an estimate of θ as well as its associated variance and combine them using Rubin's combining rules (Rubin, 2004) to obtain $\hat{\theta}$ and a 95% confidence interval (CI).

5. Report the bias of $\hat{\theta}$, the CI interval length and whether or not the CI covered the true value (Van Buuren, 2012, Section 2.5.2).

Repeat Steps 2-5 S times to obtain the empirical coverage rate. By varying the full data model and the missing data mechanism, in Step 2, we can control the two paths that influence the effectiveness of the MI procedures.

4.2 Simulation Study

In regression settings an outcome Y can depend on many explanatory variables $X_1, \ldots, X_J$ some of which can be costly to measure. As such, it is common that while the outcome Y is measured for all variables, some instances of the design matrix $\mathbf{X}$ are missing. As design elements are frequently collected prior to the outcome it is desirable to allow the imputation model access only to the design entries. To facilitate this, the missingness mechanism studied here does not allow for the missingness to depend on the outcome Y.

In this situation complete case (or listwise deletion) provides an unbiased estimate of the regression coefficients, however, the reduced sample size often leads to large standard errors and confidence intervals. When the number of explanatory variables J is of moderate size, the probability of having enough complete cases to estimate the regression coefficients is low. In this setting using a MI procedure is paramount and leads to a major reduction in the standard errors however, this can induce a slight bias. White and Carlin (2010) show

Correlation (ρ)	Missingness Coefficient (MC)
0.2	0.3
0.35	0.4
0.5	0.5
0.65	0.6

Table 1: Simulation Study configurations.

through a large simulation study that the increase in bias often time leads to a decreased empirical coverage rate for both MAR and MNAR data sets.

For our simulation study we set $J = 40$, $N = 1000$, and consider X_j that include both continuous and discrete variables in order to demonstrate the versatility of the copula approach without specifying any of the marginal distributions. This is exactly the scenario we described above; the probability of enough complete cases existing to estimate the regression coefficients is effectively 0.

The distributions we consider for the elements of the design matrix are Gaussian, Bernoulli, Poisson and ordinal. To make imputation feasible we require the variables to be correlated. To generate correlated variables we first construct a matrix of correlated Gaussian random variables and then transform the variables to have the appropriate marginals. For example, to generate a pair of correlated Poisson random variables A and B with mean λ we construct $(Z_1, Z_2) \sim \mathcal{N}(0, \Sigma)$ where $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = \rho$ and set $A = F^{-1}_{\text{Pois},4}(F_{\mathcal{N}}(Z_1))$ and $B = F^{-1}_{\text{Pois},4}(F_{\mathcal{N}}(Z_2))$. The data generating process thus leads to the following marginal distributions for the entries in $\mathbf{X}$: for $j = 1, \ldots, 10$

$$\begin{aligned} X_j &\sim \mathcal{N}(0, \sigma_j^2) & X_{j+10} &\sim \text{Bern}(p_j) \\ X_{j+20} &\sim \text{Pois}(\lambda_j) & X_{j+30} &\sim \text{ordinal}(0, 1) \\ \mathbf{X} & & &= (X_1, \ldots, X_{40}) \end{aligned}$$

$$Y \sim \mathcal{N}\left(\sum_{i=1}^{40} X_i, 1\right)$$

where $\sigma_j = 1 + (j-1)/9$, $\lambda_j = 0.2 + 2(j-1)/90$ and $p_j = 2 + 3(j-1)/9$. Both the amount of missingness (MC) and correlation (ρ) between the different variables is varied according to the specified values given in Table 1.

We consider two missing data mechanisms for $\mathbf{X}$, one that produces MAR data sets and one that produces MNAR data sets, details of which are given in Appendices A and B. Even though the MI procedures we considered are only valid under the MAR assumption, we believe it is important to consider how each method performs when this assumption is violated.

4.3 Results

We performed 1,000 simulations under each of the possible combinations of the correlation and missingness coefficient, as detailed in Table 1, under both MAR and MNAR missing data mechanisms. For MICE we specified the correct marginal distributions (for example ordered logit model for the ordinal variables) and for Amelia we used the appropriate variable transformation. In contrast we did not need to specify any distributions/transformations for the copula method. Under each procedure we created 20 completed data sets which were used to estimate the regression coefficients as well as their variances and a 95% CI. Throughout the simulation, the Amelia II software crashed a number of times, as detailed in Table D.1 in Appendix D. Thus our results for Amelia are only on a subset of the 36,000 simulations. None of the simulations had enough complete cases to estimate the regression coefficients using listwise deletion.

From Figure 1 and 2 we see that the results from all three methods were comparable with no clear procedure consistently outperforming the others. Overall the copula method had an average coverage rate of 93.2% which was much higher than that of MICE, 87.1%, and Amelia, 83%. Both the copula and MICE methods had overall bias of 0.17 while Amelia was slightly more biased at 0.25.

The copula imputations were obtained using 10,000 iterations from Hoff (2010) package whose convergence was checked on a subset of simulations. The lag-10 autocorrelation for the thinned chained is less than 0.18 in absolute value for each of the elements of the latent correlation matrix and the effective sample size was always above 200 (97.6% of the entries were above 500). The copula method had the lowest bias, highest coverage rate and the longest interval length. Even though the semi-parametric estimation procedure did not require specification of the marginals, any data transformations, or tuning it still outperformed the other two procedures.

Since the MICE procedure is iterative, one must check that the model parameters fully explore the parameter space. Unlike the Bayesian copula method, there are no explicit convergence criteria one can track – however, we performed a visual check that revealed no abnormalities and also ran each MICE chain for 20 iterations as recommended in Van Buuren and Groothuis-Oudshoorn (2011). The MICE method performed almost as well as the copula method, but had slightly lower coverage rate. It also had the lowest average bias for the normal and Poisson variables. Again, however, these results are contingent on specifying the correct conditional distribution which can often be challenging.

Amelia had the lowest coverage and highest bias both on average and in most scenarios that we considered. It had the lowest average interval length of 1.23, which shows that it was systematically underestimating the variance – leading to the low coverage rates.

Figure 1 shows that the absolute mean bias and the interval length increases as a function of the proportion of missing values, under both missing data mechanisms. The coverage also starts to worsen mainly due to the increased bias. One notable exception was the good coverage properties of the copula approach for the regression parameter of the ordinal variables, both Amelia and MICE here undercovered the true value. As this marginal structure is very frequently encountered in social science applications. The copula method also has the lowest bias and interval length for the binomial regression coefficient and yet still has the best coverage properties.

Surprisingly there seems to be little variation in the bias and the interval length as a function of the correlation, as is shown in Figure 2.

Breaking the MAR assumption did not lead to drastically worse results. There was

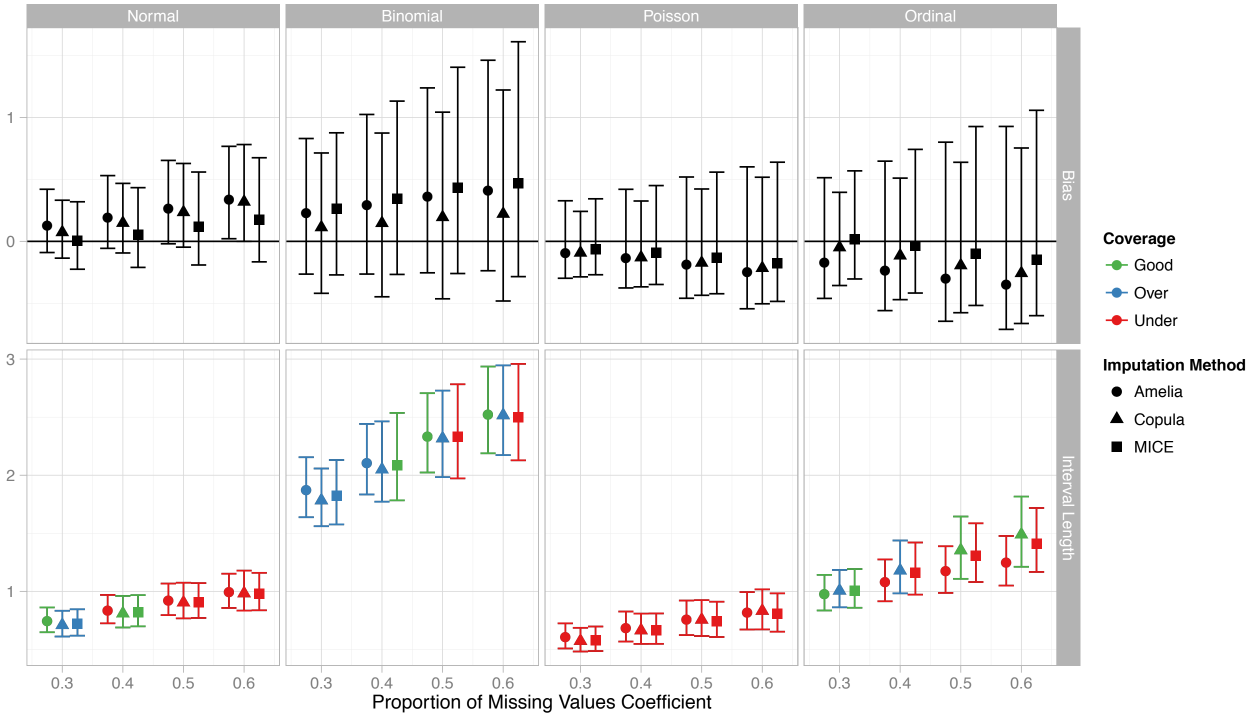

Figure 1: Simulation study results for the MAAR missing data mechanisms. The errorbars we obtained using the 2.5% and 97.5% quantiles of the simulation distribution. The colors represent good coverage (green – between 92.5% and 96.5%), over coverage (blue – above 96.5%) and under coverage (red – below 92.5%).

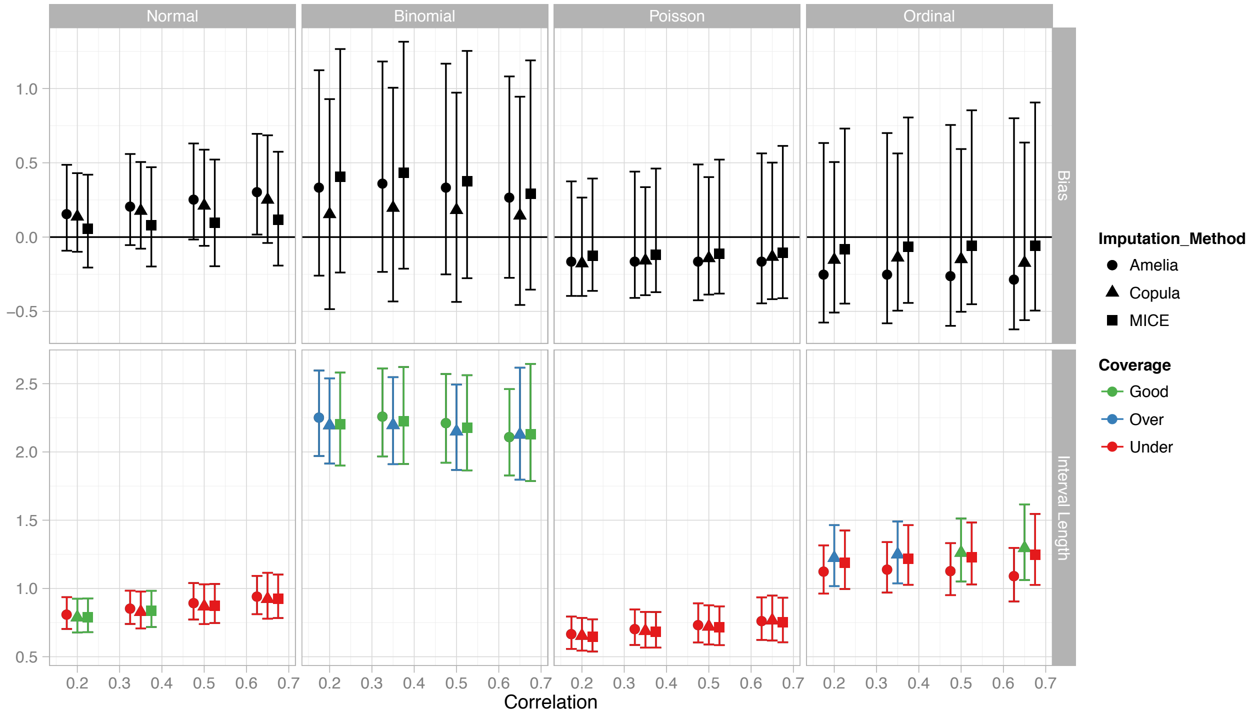

Figure 2: Simulation study results for the MAAR missing data mechanisms. The errorbars we obtained using the 2.5% and 97.5% quantiles of the simulation distribution. The colors represent good coverage (green – between 92.5% and 96.5%), over coverage (blue – above 96.5%) and under coverage (red – below 92.5%).

a decrease of about 3% in the coverage of all three methods and a slight decrease in the average bias. This shows that the methods are somewhat robust to violations of MAR assumption when it is not too severe. Figures C.1 and C.2 in the Appendix C show the results of the simulations when the MAR assumption is violated.

5 Application Study

In this section, we provide comparison of the three imputation methods using an application from political science. The application shows how copula methods can be used to impute a large data set with a variety of variable types.

5.1 Inequality and Democratic Support

As we have elaborated above, imputation methods are still underused, especially in the social sciences. There is, however, some visible progress. One example where scholars have taken advantage of one of the imputation methods currently available is "Economic Inequality and Democratic Support" by Krieckhaus et al. (2014) published in the *Journal of Politics*. Krieckhaus et al. (2014) explore whether the support for democracy within countries is affected by the level of inequality. The authors combine country level variables (such as inequality) with individual level survey data from 40 democracies around the world. For multiple countries several survey waves are included, resulting in 57 country-years and a total of 77,642 observations (Krieckhaus et al., 2014, 144). For this replication exercise we replicate *Model 1* in *Table 1* in Krieckhaus et al. (2014). The dependent variable is a "13-point additive index (ranging from 0 to 12) of democratic support", which the authors treat as a continuous variable (Krieckhaus et al., 2014, 144). The main independent variables of interest are *Inequality* at the country level, and an ordinal *income* scale at the individual level (ranging from 1 to 10). Additionally, the authors control for *Age*, *Gender*, *Institutional Confidence*, *Interest in Politics*, *Interpersonal Trust*, *Education*, *Prior Regime Evaluation*, and *Leftist Ideology* all drawn from the *World Values Survey* (World Values Survey, 2012). As in the original article, all individual level variables are demeaned "using group-mean centering" after the imputation (Krieckhaus et al., 2014, 145). The data are analyzed using a random-coefficients model.

Table 2: Share of Missingness in Variables of Interest

Democracy Support	Inequality	Income	Age
19.9	1.8	12.9	0.2
Gender	Institutional Confidence	Interest in Politics	Interpersonal Trust
0.1	11.7	2.5	3.7
Education	Leftist Ideology	Prior Regime Evaluation	
3.9	18.5	21.3	

Most importantly for the purpose of this study, the original data suffers from a relatively high number of missing observations. Table 2 shows the share of missing observations for variables included in the replication exercise. As one can see, a number of variables have a large share of missing observations. If instead of multiple imputation, the authors would have engaged in listwise deletion with respect to the missing data, the number of observation in the regression model would have been approximately halved. Instead, Krieckhaus et al. (2014) use Amelia II (Honaker et al., 2013) to multiple impute five data sets which they analyze. Estimates are then combined using Rubin's rule.

This is an excellent setting for our comparison of multiple imputation techniques. The number of missing observations is quite large and we have a number of different types of variables, continuous, binary, as well as ordinal. We create 20 multiple imputed data sets using each of the imputation techniques: Amelia II, MICE, and Copula. We then re-estimate *Model 1* in *Table 1* in Krieckhaus et al. (2014, 147) and combine the estimation results for each method's multiple imputed data sets via Rubin's rule.

For Amelia II we specify the type of each variable and then generate 20 imputed data sets using the full original data. Similarly, we declare each variable's type for MICE and estimate the default model for each of these types. We use all variables except the one to be imputed as independent variables in the chained equations. Again we create 20 multiple imputed data sets and set the maximum number of iterations to 20. Lastly, we use our preferred method, imputation via the semi-parametric Gaussian copula, to generate 20 imputed data sets. We run the MCMC chain for 2100 iterations and randomly draw 20 data sets from the posterior.

Figure 3 shows the coefficient estimates and 95% confidence intervals for the replicated model based on each of the imputation techniques, as well as when list-wise deletion is used. First, for the majority of variables included in the regression model the results are relatively similar across the different imputation techniques and even for the list-wise deletion. In fact, for the two main variables of interest, inequality and income, the results based on different imputation techniques are virtually the same.

On the other hand, there are several significant differences. First, the effect of gender is essentially zero according to the models estimated on the copula imputed data. Based on the data imputed using MICE or Amelia II females have higher ratings of democracy satisfaction (though the confidence intervals just cover zero). According to the non-imputed data, the effect of gender is quite strong and precisely estimated. Similarly, based on the data imputed with Copula method, the estimated associations of *Education*, *Interest in Politics*, and *Prior Regime Evaluation* with the dependent variable of *Democracy Satisfaction* are all weaker, compared to the other methods (and the non-imputed data), though the confidence intervals overlap. On the other hand, the associations of *Leftist Ideology* and *Institutional Confidence* with *Democracy Satisfaction* are estimated to be much stronger based on the copula imputed data, compared to the other imputation methods. Here the estimated coefficients based on the different imputation techniques are quite different and the confidence intervals do not overlap.

It is interesting to note, that, except for one variable (*Interpersonal Trust*), whenever the estimated coefficient for the copula imputed data differs from the coefficients based on the other imputation methods, it is in the opposite direction of the difference to the list-wise deletion coefficient. This is especially easy to see for the *Gender* and *Leftist Ideology* variables, where the effect is strongest (weakest) according to the model estimated on the list-wise deleted data and weakest (strongest) for the copula based models.

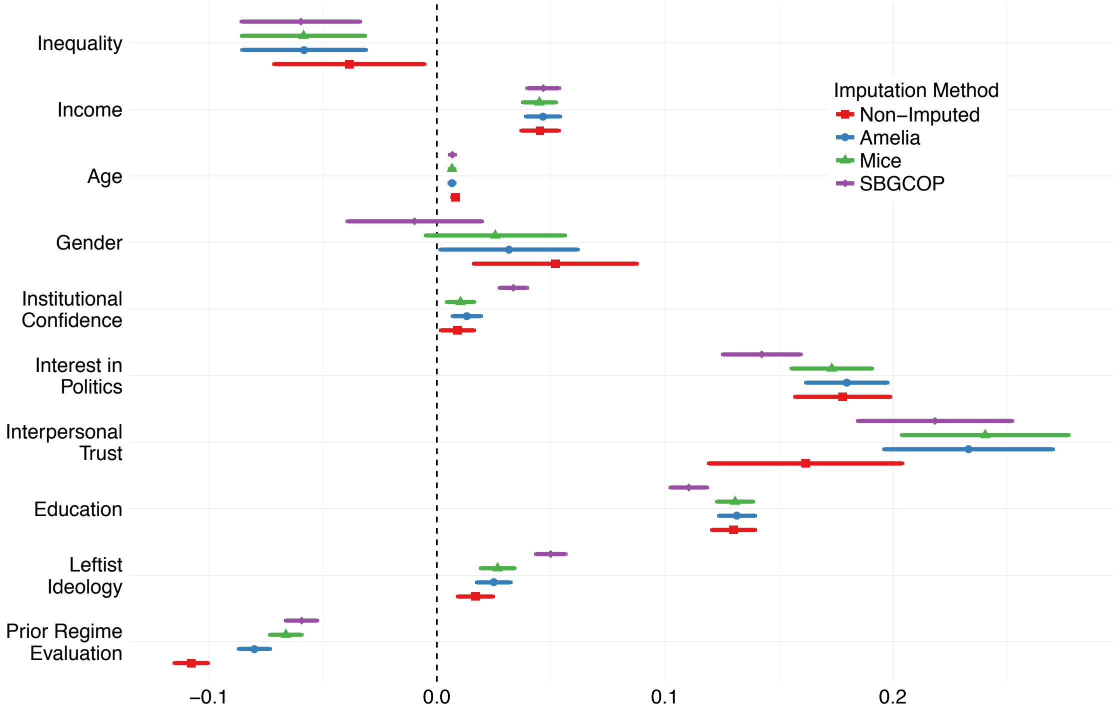

Figure 3: Coefficient estimates and confidence intervals for *Model 1* in *Table 1* in Krieckhaus et al. (2014) based on three imputation techniques and list-wise deletion

6 Conclusion

What practical lessons can we learn about how to deal with missing data? Despite the fact that missing data is ubiquitous, still, too few authors beyond the statistics community make use of statistical methods (King et al., 2001, Honaker and King, 2010, Van Buuren and Groothuis-Oudshoorn, 2011, Kropko et al., 2014) to deal with this problem. In this article, we re-emphasize the importance of dealing with missing data and present a copula based approach, developed by Hoff (2007), that is elegant and requires no pre-specification of the data. With the rank based approach introduced by Hoff (2007), copulas can also be used to impute binary, ordinal, and continuous variables and it generally performs as well as or even better than either `Amelia II` or `MICE`.

Throughout our simulation and the application, the three imputation methods perform relatively similarly, but with subtle differences. MICE, however, requires specification of conditional distributions while the copula method does not. Moreover, recent theoretical results for MICE suggest that good performance heavily relies on being approximately correct in the choice of conditionals (Li et al., 2012). Theoretical guarantees for good behavior of copula methods are available. In particular, information bounds for rank-based estimators are the same as the information bounds for estimators based on the full (scale and rank) data (Hoff et al., 2014). Under MAR and MCAR we inherit all the properties of the full data and by introducing structure to the imputation we are likely to have good behavior even under MNAR.

One of the advantages of using the semiparametric copula approach to impute large data sets is the relatively limited memory and computational requirements. As indicated by Graham (2009) the disadvantages of EM approaches are especially large when imputing databases with many variables or applications of "big data". While it can be computationally less expensive, MICE suffers when the number of variables increases as the correct choice of specification for each of the conditionals becomes increasingly unlikely. In contrast, the use of the semiparametric copula makes it possible to impute even large database in a relative timely manner, with limited computing resources, and no pre-specification of the data. Moreover, using the copula model to multiple impute missing values provides some of the advantages (such as a proper posterior distribution of the data) but is less burdensome on scholars than imputing values in a fully Bayesian approach (Erler et al., 2016).

Finally, the copula approach is quite flexible and can be employed at different stages of the analysis process. First, it can be used to generate a single estimate of the missing data or the mean of a large number of draws, which is exactly what might be needed in some situations. Second, per the recommendation of Rubin, it can be used to construct multiple databases. These results can then be averaged, thereby accounting for a portion of the uncertainty in the imputed values. While this is not based on E-M, it does allow one to have a principled set of databases that are absent missing data. As with `Amelia II`, the copula imputations can be analyzed separately and the results combined using either `mitools` or `Zelig` (Imai et al., 2008) in $\mathcal{R}$. Thus, the copula approach to missing data can be explicitly integrated into the modeling and analysis of observational data in a simplistic, organic fashion that is computationally efficient.

A Missing at Random

Since we are interested in evaluating the frequentest properties of the different MI procedures we need to use a missing data mechanism that always produces MAR data, this is known as the Missing Always at Random (MAAR) assumption (Mealli and Rubin, 2015). Below we describe the MAAR missing data mechanism that we used.

1. Given a fully observed data set $\mathbf{X}$ randomly select four variables, one from each of the four classes, that will be fully observed; without loss of generality relabel them X_1, X_{11}, X_{21} and X_{31}.

2. Randomly select four variables from the remaining thirty six, one from each of the four classes, that will have a 5-6% missingness; without loss of generality relabel them X_2, X_{12}, X_{22} and X_{32}. The probability that the i^{th} observation for each variable is missing is based on a logistic regression on the fully observed variables, X_1, X_{11}, X_{21} and X_{31}, adjusted so that the mean number of missing variables is between 5-6%. The missingness indicators are then sampled from independent Bernoulli random variables with the appropriate probabilities. Let $\mathbf{X}^{(1)} = (X_1, X_2, X_{11}, X_{12}, X_{21}, X_{22}, X_{31}, X_{32})$ and $\mathbf{X}^{(1)}_{\text{cc}}$ be the complete cases after removing the any rows that have missing values.

3. For the remaining thirty two variables the probability of the i^{th} observation missing is based on a logistic regression on the fully observed $\mathbf{X}^{(1)}_{\text{cc}}$ adjusted so that the mean number of missing variables is equal to the Missingness Coefficient (MC) (see Table 1 for the range of values that we considered). The missingness indicators are again sampled from independent Bernoulli random variables with the appropriate

probabilities. If the i^{th} row of $\mathbf{X}^{(1)}$ has been removed in $\mathbf{X}_{cc}^{(1)}$ then that row is always observed for the thirty two variables.

Since we have eight fully/almost fully observed variables the overall number of missing values produced by this missingness mechanism is slightly lower than the values of MC given in Table 1.

B Missing not at Random

Creating a MNAR data set is easier than a MAR data set. The below algorithms describes our missing always not at random (MANAR) mechanism which, with extremely high probability, produces a MNAR data set.

1. Given a fully observed data set $\mathbf{X}$ randomly select four variables, one from each of the four classes, that will be fully observed; without loss of generality relabel them X_1, X_{11}, X_{21} and X_{31}.

2. Randomly select four variables from the remaining thirty six, one from each of the four classes, that will have a small amount of missingness; without loss of generality relabel them X_2, X_{12}, X_{22} and X_{32}. The probability that the i^{th} observation is missing

is given by,

$$P(R_2 = 1|\mathbf{X}) = 1_{X_2>0}p_{MC},$$

$$P(R_{12} = 1|\mathbf{X}) = 1_{X_{12}=0}p_{MC},$$

$$P(R_{22} = 1|\mathbf{X}) = 1_{X_{22}>3}p_{MC},$$

$$P(R_{32} = 1|\mathbf{X}) = 1_{X_{32}=3}p_{MC},$$

where the value of p_{MC} is given by the MC in Table 1.

3. For the remaining thirty two variables the probability of the i^{th} observation missing is based on a logistic regression on $\mathbf{X}^{(1)}$ adjusted so that the mean number of missing variables is equal to the MC (see Table 1). The missingness indicators are again sampled from independent Bernoulli random variables with the appropriate probabilities. In contrast to the MAAR mechanism if the i^{th} row of $\mathbf{X}^{(1)}$ has missing values then other variables in that row can still be missing.

C Plots of MNAR Simulation Results

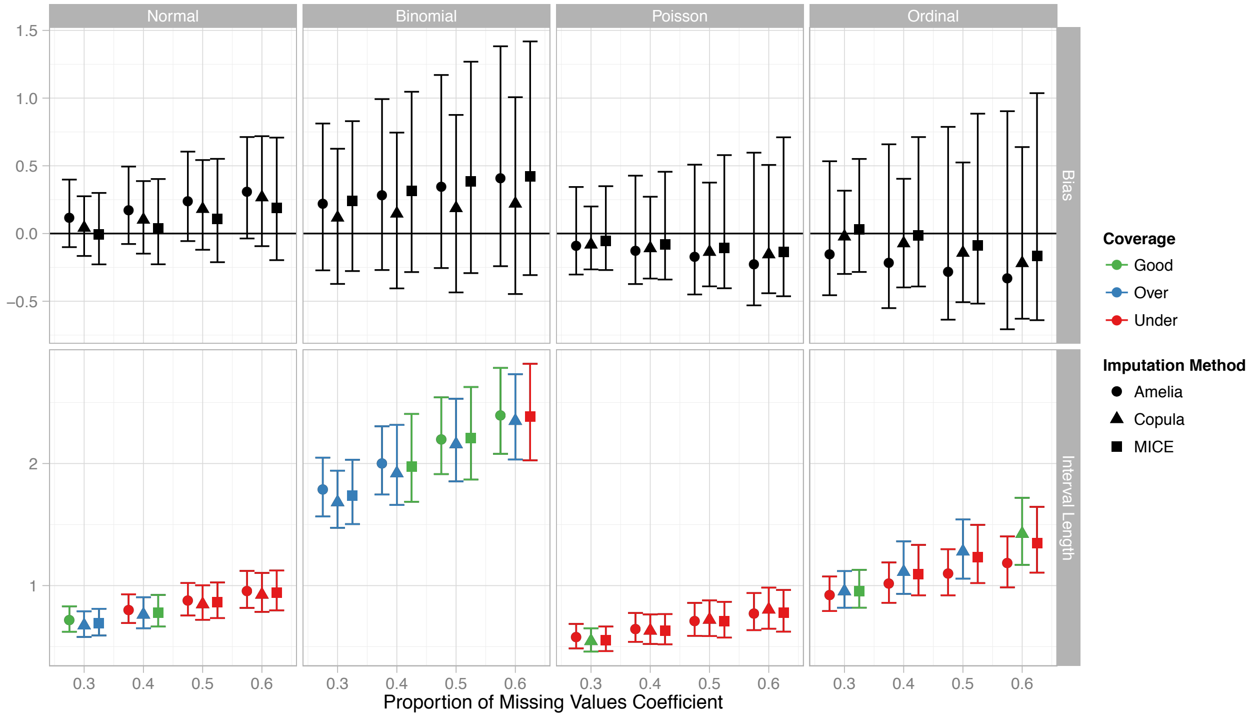

Figure C.1: Simulation study results for the MANAR missing data mechanisms. The errorbars we obtained using the 2.5% and 97.5% quantiles of the simulation distribution. The colors represent good coverage (green – between 92.5% and 96.5%), over coverage (blue – above 96.5%) and under coverage (red – below 92.5%).

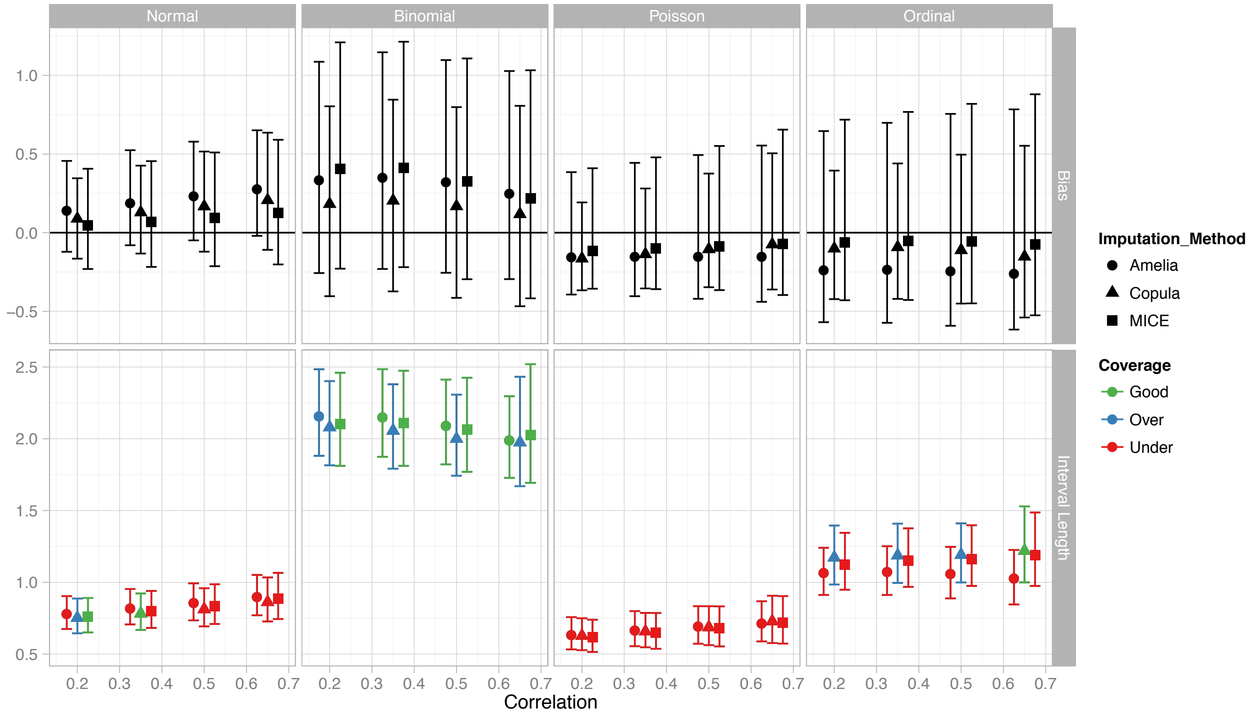

Figure C.2: Simulation study results for the MANAR missing data mechanisms. The errorbars we obtained using the 2.5% and 97.5% quantiles of the simulation distribution. The colors represent good coverage (green – between 92.5% and 96.5%), over coverage (blue – above 96.5%) and under coverage (red – below 92.5%).

D Number of Simulations for which Amelia crashed

		Correlation			
		0.2	0.35	0.5	0.65
	0.3	2	0	0	7
Share of	0.4	93	16	8	0
Missingness	0.5	285	138	37	13
	0.6	485	305	159	72

Table D.1: The number of Amelia crashes out of the 1000 simulations under each of the possible scenarios.

References

Blackwell, M., J. Honaker, and G. King (2015). Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data. *Sociological Methods and Research In Press*.

Chen, X., Y. Fan, and V. Tsyrennikov (2006). Efficient Estimation of Semiparametric Multivariate Copula Models. *Journal of the American Statistical Association 101*(475), 1228–1240.

Dempster, A. P., N. M. Laird, and D. 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.

Di Lascio, F., S. Giannerini, and A. Reale (2015). Exploring copulas for the imputation of complex dependent data. *Statistical Methods & Application 24*(1), 159–174.

Erler, N. S., D. Rizopoulos, J. v. Rosmalen, V. W. Jaddoe, O. H. Franco, and E. M. Lesaffre (2016). Dealing with Missing Covariates in Epidemiologic Studies: A Comparison Between Multiple Imputation and a Full Bayesian Approach. *Statistics in Medicine*.

Goodrich, B., J. Kropko, A. Gelman, and J. Hill (2012). mi: Iterative Multiple Imputation from Conditional Distributions. R package.

Graham, J. W. (2009). Missing Data Analysis: Making it Work in the Real World. *Annual Review of Psychology 60*(1), 549–576.

Hoff, P. (2010). *sbgcop: Semiparametric Bayesian Gaussian Copula Estimation and Imputation.* R package version 0.975. `https://CRAN.R-project.org/package=sbgcop`.

Hoff, P. D. (2007). Extending the Rank Likelihood for Semiparametric Copula Estimation. *Annals of Applied Statistics 1*(1), 265–283.

Hoff, P. D., X. Niu, and J. A. Wellner (2014). Information Bounds for Gaussian Copulas. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability 20*(2), 604.

Honaker, J. and G. King (2010, April). What to do About Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science 54*(2), 561–581.

Honaker, J., G. King, and M. Blackwell (2013). AMELIA II: A Program for Missing Data.

Horton, N. J. and K. P. Kleinman (2007, February). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician 61*(1), 79–90.

Imai, K., G. King, and O. Lau (2008). Toward A Common Framework for Statistical Analysis and Development. *Journal of Computational and Graphical Statistics 17*(4), 892–913.

Käärik, E. and M. Käärik (2009). Modeling dropouts by conditional distribution, a copula-based approach. *Journal of Statistical Planning and Inference 139*, 3830–3835.

King, G., J. Honaker, A. Joseph, and K. Scheve (2001, March). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Politial Science Review 95*(1), 49–69.

Klaassen, C. A., J. A. Wellner, et al. (1997). Efficient Estimation in the Bivariate Normal Copula Model: Normal Margins are Least Favourable. *Bernoulli 3*(1), 55–77.

Krieckhaus, J., B. Son, N. Bellinger, and J. Wells (2014). Economic Inequality and Democratic Support. *The Journal of Politics 76*(1), 139–151.

Kropko, J., B. Goodrich, A. Gelman, and J. Hill (2014). Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches. *Political Analysis 22*(4), 497–519.

Li, F., Y. Yu, and D. B. Rubin (2012). Imputing Missing Data by Fully Conditional Models:

Some Cautionary Examples and Guidelines. *Duke University Department of Statistical Science Discussion Paper 1124*.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (second ed.). New York: Wiley.

Liu, J., A. Gelman, J. Hill, Y.-S. Su, and J. Kropko (2013). On the Stationary Distribution of Iterative Imputations. *Biometrika 101*(1), 155–173.

Mealli, F. and D. B. Rubin (2015). Clarifying Missing at Random and Related Definitions, and Implications when Coupled with Exchangeability. *Biometrika 102*(4), 995–1000.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2014). *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman and Hall/CRC.

Murray, J. S. (2013). Some Recent Advances in Non- and Semiparametric Bayesian Modeling with Copulas, Mixtures, and Latent Variables. Dissertation. Department of Statistical Science Duke University. `http://dukespace.lib.duke.edu/dspace/handle/10161/8253`.

Murray, J. S., D. B. Dunson, L. Carin, and J. E. Lucas (2013). Bayesian Gaussian Copula Factor Models for Mixed Data. *Journal of the American Statistical Association 108*(502), 656–665.

Pettitt, A. (1982). Inference for the Linear Model using a Likelihood Based on Ranks. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(2), 234–243.

Pitt, M., D. Chan, and R. Kohn (2006). Efficient Bayesian Inference for Gaussian Copula Regression Models. *Biometrika 93*(3), 537–554.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1987a). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons.

Rubin, D. B. (1987b). *Statistical Analysis with Missing Data.* John Wiley & Sons.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American statistical Association 91*(434), 473–489.

Rubin, D. B. (2004). *Multiple imputation for Nonresponse in Surveys*, Volume 81. John Wiley & Sons.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data.* New York, NY: Chapman & Hall.

Sklar, A. (1959). Fonctions de Répartition à N Dimensions et Leur Marges. *Publications de l'Institut Statistique de l'Université Paris 8*, 229–231.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Boca Raton, FL: Chapman & Hall/CRC Press.

Van Buuren, S. and K. Groothuis-Oudshoorn (1999). Flexible Multivariate Imputation by MICE. *Leiden, The Netherlands: TNO Prevention Center*, 1–20.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software 45*(3), 1–67.

Vuolo, M. (2015). Copula Models for Sociology. *Sociological Methods and Research In Press*, 1–45.

White, I. R. and J. B. Carlin (2010). Bias and Efficiency of Multiple Imputation Compared with Complete-Case Analysis for Missing Covariate Values. *Statistics in Medicine 29*(28), 2920–2931.

World Values Survey (2012). 1981-2008 Integrated Questionnaire.

Yucel, R. M. (2011). State of the Multiple Imputation Software. *Journal of Statistical Software 45*(1), 1 – 7.

Zhu, J. and T. E. Raghunathan (2015). Convergence Properties of a Sequential Regression Multiple Imputation Algorithm. *Journal of the American Statistical Association 110*(511), 1112–1124.