

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

October 4, 2018

*Accepted for publication at *Sociological Methods & Research*. 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); Shahryar Minhas is an Assistant Professor, Department of Political Science, Michigan State University, East Lansing, MI, 48824 (email: minhassh@msu.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); 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 empirical research, especially in the social sciences. Despite multiple approaches to dealing adequately with missing data, many scholars still fail to address this vital issue. In this paper, we present a simple-to-use method for generating multiple imputations using a Gaussian copula. The Gaussian copula for multiple imputation (Hoff, 2007) allows scholars to attain estimation results that have good coverage and small bias. The use of copulas to model the dependence among variables will enable researchers to construct valid joint distributions of the data, even without knowledge of the actual underlying marginal distributions. Multiple imputations are then generated by drawing observations from the resulting posterior joint distribution and replacing the missing values. Using simulated and observational data from published social science research, we compare imputation via Gaussian copulas with two other widely used imputation methods: MICE and Amelia II. Our results suggest that the Gaussian copula approach has a slightly smaller bias, higher coverage rates, and narrower confidence intervals compared to the other methods. This is especially true when the variables with missing data are not normally distributed. These results, combined with theoretical guarantees and ease-of-use suggest that the approach examined provides an attractive alternative for applied researchers undertaking multiple imputations.

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 is widely known to lead to biased results, and inflated (or deflated) variance estimates (Rubin, 1976, King et al., 2001, White and Carlin, 2010, Molenberghs et al., 2014). Even though most statistical software platforms provides methods that adequately handle missing data (the most popular of these is multiple imputations (MI)), they are often ignored by applied researchers.¹

In Figure 1, we illustrate the number of articles published in five top sociology and political science journals since 1990 that contain “multiple imputations” in the body of the paper.² Our survey of the literature shows the rapid growth of the use of multiple imputations in the social sciences. Nevertheless, as missing data is a feature of almost any observational data set, the annual counts of articles mentioning multiple imputations per year still point to significant underutilization of this method in the social sciences. This may be due to a lack of understanding of the benefits (and assumptions) of common

¹Principled approaches to missing data have existed for over three decades. First formalized by Rubin (1976), the number of readily available statistical softwares to deal with missing data has rapidly grown since the 1990s (e.g. King et al., 2001, Honaker and King, 2010, Van Buuren and Groothuis-Oudshoorn, 2011, Kropko et al., 2014). Further, see the special issue on the *State of Multiple Imputation Software* in the *Journal of Statistical Software* in 2011 (Yucel, 2011).

²The five journals we reviewed from sociology are *Annual Review of Sociology*, *American Sociological Review*, *American Journal of Sociology*, *Sociological Methodology*, and *Sociological Methods & Research*. In political science we examined the *American Political Science Review*, *American Journal of Political Science*, *Political Analysis*, *British Journal of Political Science*, and the *Journal of Politics*.

imputation methods.

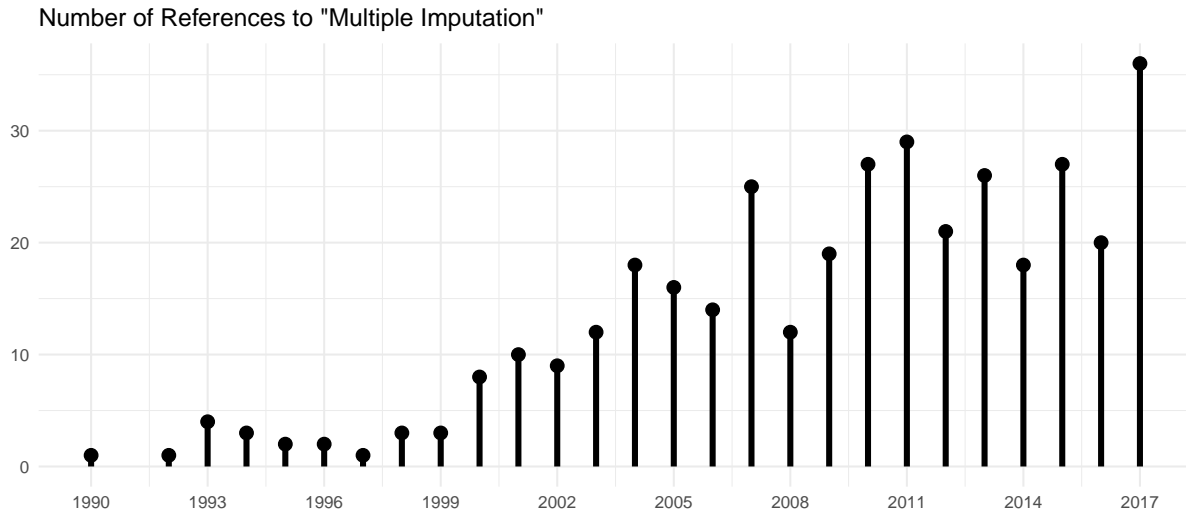


Figure 1: Number of references to “multiple imputation” in articles from five top sociology and political science journals since 1990.

This article has two aims. First, we introduce applied researchers in the social sciences to a specific copula method for imputation and discuss its advantages over other methods. The method discussed is easy to implement using the `sbgcop` package (Hoff, 2010) in \mathcal{R} (R Development Core Team, 2004) ³ and has theoretical properties that make it attractive. Second, we conduct a systematic evaluation and comparison of the copula method to two commonly used imputation software packages (MICE Van Buuren and Groothuis-Oudshoorn (2011) and AMELIA II (Honaker et al., 2012)) in sociology and political science.

Copulas are often used for the estimation of dependency between variables and are particularly useful in the generation of imputations as they allow for the construction of

³For inexperienced users, our `gcImp` (<https://github.com/bojinov/gcImp>) package provides a simple interface for generating imputations using `sbgcop`.

valid joint distributions of the data, even if the researcher has little knowledge about the actual joint distribution of the variables. Given the joint distribution of the data, we can generate imputations by sampling from the conditional distribution of the missing data given the observed data.

We highlight a semi-parametric Gaussian copula approach to missing data imputation. The Gaussian copula is one particular way of constructing a joint distribution from which missing values can be easily drawn. The method was initially developed by Hoff (2007) to estimate empirical models on multivariate data.

In particular, the Gaussian copula defines the dependence among the distributions of a set of variables which may contain missing values. These variables can include normal, ordinal, and binary variables. Rather than using the distributions themselves, a rank likelihood approximation is used. As a result, the technique does not require the specification of marginal or conditional distributions. This is in stark contrast to other imputation methods using copulas that either require knowledge of the marginals or correlation structure (Käärik, 2006, Käärik and Käärik, 2009, Robbins et al., 2013) or target different copula parameters via pseudolikelihood methods (Di Lascio et al., 2015). The proposed approach allows applied researchers to undertake imputations of their data without relying on pre-specification or ad-hoc decisions.

The potential use of copulas for multiple imputation applications has not been thoroughly discussed within the social sciences. The copula methods we describe are easy to use and are more likely to provide a good representation of the joint distribution of the data than existing methods. Moreover, provided the Markov Chain Monte Carlo (MCMC)

converges, the output from the copula model represents a valid posterior density. Simply put, this means that we have theoretical guarantees about the posterior distribution from which the imputations are generated that other methods can not provide. Based on an extensive simulation exercise, we show that the method presented here is generally at least as accurate as other commonly used methods—it is often better. It also provides better uncertainty estimates for the imputations. Lastly, as is shown in Bojinov et al. (2017), the copula method can also be used to test some of the underlying assumptions about the appropriateness of imputations for a given data set.

2 Common Approaches to Multiple Imputation

The standard techniques employed to deal with missing data require an assumption regarding the missing data pattern; these were first formalized in Rubin (1976).⁴ To briefly summarize these terms, 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 (Marini et al., 1980). 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.

⁴Little and Rubin (2002) provide a more up to date treatment and Mealli and Rubin (2015) an in-depth discussion on the different missing data mechanisms.

The most basic methods, such as listwise deletion, generally lead to biased regression coefficients if the missingness process is not **MCAR** (Graham, 2009). To achieve valid inference under the Bayesian and likelihood paradigms, while ignoring the missing data mechanism, we require the weaker **MAR** assumption.⁵

The most common appropriate approach to dealing with missing data is multiple imputation (MI), which refers to any method that replaces the set of missing values with various plausible values, thus obtaining m completed data sets (Rubin, 1996). Rubin (1987) initially suggested creating five imputations, but more recently authors recommended using closer to twenty imputations (Van Buuren, 2012).⁶ The completed data sets are then separately analyzed using the standard full data techniques and the resulting quantities of interest from each data set are combined to obtain an overall, average estimate as well as its associated variance.

Before moving to introduce the copula method below, we briefly outline two important methods for generating multiple imputations here.

MI with EM This approach uses iterative expectation maximization (EM) to create complete data sets based on assuming a particular joint distribution. A widely used method for imputation in the social sciences is the **Amelia II** \mathcal{R} package by Honaker

⁵A further assumption of parameter distinctness—the parameter governing the data and the parameter governing the missingness mechanism are *a priori* independent—is required to ensure that valid statistical inference whenever the data are **MAR** or **MCAR**. See Little and Rubin (2002) for more details on this assumption.

⁶This was based on examining large sample relative efficiency when using a finite number of proper imputations compared to an infinite number, from a Bayesian Gaussian model. In practice, non-normal data combined with non-Bayesian methods can lead to a decrease in the relative efficiency.

and King (2010) In *Amelia II*, the joint distribution of the data is modeled as a multivariate normal distribution. *Amelia II* provides an implementation of the EM approach by the use of bootstrapping to derive solutions quickly. One of the disadvantages of imputation via EM is that for large data sets with significant amounts of missing data, it is computationally intensive. This is a trait of EM algorithms in general, as the rate of convergence is proportional to the amount of missing information in the model. Moreover, it is often unclear to what degree modeling the joint distribution of the data as a multivariate normal distribution is appropriate, especially since the data may include binomial and ordinal variables.

Conditional Approaches to Multiple Imputation An alternative method is to model

each variable’s imputation via its conditional distribution based on all other variables in the data. One such approach is developed in Multiple Imputation via Chained Equations (MICE) (Van Buuren, 2012), another was developed as the *MI* package in \mathcal{R} (Goodrich et al., 2012). Imputations for fully conditional specification (FCS) methods, such as MICE or MI, are created 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 only under certain conditions do the individual conditional models define a valid joint distribution. This often leads to pathologies in the convergence of the algorithms (Li et al., 2012, Chen and Ip, 2015). For example, if $Y|X$ is specified to be an Exponential random variable with rate X and $X|Y$ is specified to be an Exponential random variable with rate Y , it

is well known that no joint distribution exists and sequentially sampling from these two distributions generates draws that tend to infinity (Casella and George, 1992). More strikingly, Example 1 of Li et al. (2012) demonstrates that even when all the conditionals are normal, the order in which the variables are updated in MICE can determine whether the chain will converge to a stationary distribution.

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 also means the imputation model for each variable in the data has to be correctly specified, which 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).

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 (such as the example in the previous section).⁷ When a valid joint distribution does not exist, then there are no guarantees that

⁷Some theoretical results for MICE are available, but they do not allow too much misspecification in the conditional models. For example, 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 correct joint 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

the MI procedure is proper (as defined in Rubin, 2004). A natural approach to overcoming a possibly incompatible conditional specification is by specifying the joint distribution directly. For example, this is done in most EM approaches, such as `Amelia II`, by simply assuming a multivariate normal distribution. However, while an approximation, most social science data include binary and ordinal variables, and thus cannot have a multivariate normal joint distribution. As a result, this misspecification of the joint distribution is problematic. Moreover, specifying the correct joint distribution becomes increasingly complicated as the number of covariates in the model increase.

It is therefore valuable to decouple the specification of the marginal distribution of each covariate from the function that describes the joint behavior of all covariates together. One of the main advantages of using copulas for imputation is that they allow us to do exactly that. 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 p -dimensional joint distribution function with marginals F_1, \dots, F_p . Then there exists a copula C with uniform marginals such that*

$$F(x_1, \dots, x_p) = C(F_1(x_1), \dots, F_p(x_p))$$

Sklar’s theorem guarantees that the function C is unique if the marginal distributions F_1, \dots, F_p are continuous. If they are discrete, then it is unique on the cross product of the ranges of the F_j .

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

theoretical results to missingness patterns where each individual is missing at most one variable).

matrix R (Klaassen et al., 1997, Pitt et al., 2006, Chen et al., 2006, Hoff, 2007). That is, we define the Gaussian copula function as $C(\cdot|R) = \Phi_p(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p)|R)$ for $u_1, \dots, u_p \in (0, 1)^p$ where Φ is the univariate normal CDF and $\Phi_p(\cdot|R)$ is the p -dimensional CDF with correlation matrix R . This means that the joint distribution of the p variables is given by $F(x_1, \dots, x_p) = \Phi_p(\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_p(x_p))|R)$. Simply put, the univariate CDFs F_1, \dots, F_p of the individual variables are bound together as a multivariate normal CDF where R determines the correlation between the individual variables on the normal scale.

As previously noted, the specification of marginal distributions is difficult in applied settings and so of particular interest is the setting where the researcher does not need to specify the marginal distributions for F_1, \dots, F_p . In fact, one big advantage to the method discussed here is that we consider a semiparametric approach that does not require parameterizing the p marginal distributions.

In this flexible setting, the estimation procedures described below provide consistent and likely asymptotically efficient estimates of the dependence parameters in the Gaussian copula, *i.e.*, R above (Murray et al., 2013, Hoff et al., 2014). These dependence parameters directly impact the imputation of the missing data, and thus 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 of Pettitt (1982) to the copula setting.

The rank likelihood (Pettitt, 1982) is a type of marginal likelihood that bases inference on the ranks of data rather than the full data. In a univariate setting it is defined as follows: consider $z_1, \dots, z_n|\theta \sim p(z|\theta)$ be a sample from some distribution. Instead of

observing the actual values z_1, \dots, z_n , however, consider only observing the ordering of the data x_1, \dots, x_n (*i.e.* their rank). Then the rank likelihood is given by

$$L(\theta; x_1, \dots, x_n) = \int_D p(z_1, \dots, z_n | \theta) dz_1, \dots, dz_n$$

where $D = \{z_{\alpha_1} < \dots < z_{\alpha_n}\}$ and $\alpha_i = j$ if and only if z_j is the i th smallest of z_1, \dots, z_n .

Hoff (2007) extends the rank likelihood to the multivariate setting by considering the semiparametric Gaussian copula. Let $z_1, \dots, z_n | R \sim N(0, R)$, with $z_i = (z_{i1}, \dots, z_{ip})$, and let $x_{ij} = F_j^{-1}(\Phi(z_{ij}))$. That is, latent data are drawn from a multivariate normal distribution with correlation R and are transformed to the observed scale via an inverse transformation as in the definition of the Gaussian copula above. One can consider the observed data as the ranks of the unobserved latent Z s and define

$$D = \{Z \in R^{n \times p} : \max\{z_{kj} : x_{kj} < x_{ij}\} < z_{ij} < \min\{z_{kj} : x_{ij} < x_{kj}\}\}.$$

It is easy to see that all $Z \in D$ respect the order of the variables on the observed scale. Hoff (2007) shows that $P(Z \in D | R, F_1, \dots, F_p) = P(Z \in D | R)$ which in turn allows for the decomposition

$$P(X | R, F_1, \dots, F_p) = P(Z \in D | R) P(X | Z \in D, F_1, \dots, F_p).$$

The aforementioned results guarantee that inference about R can proceed simply via $P(Z \in D | R)$. This leverages the ordering of the observed values x_{1j}, \dots, x_{nj} of each variable to make inference about the parameter R without estimating the CDFs F_1, \dots, F_p .

This means that regardless of the marginal distributions of the individual variables, all we need is their ordering to facilitate the use of the Gaussian copula model to make

inferences about the dependence between these variables, *i.e.*, the correlation matrix R . 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 since full conditional distributions can be derived by conditioning on the ranks of the data alone.⁸

Let us paraphrase and summarize the method in less technical terms. Assume we have two vectors Z_1 and Z_2 which come from a bivariate normal distribution with correlation R . We observe $X_i = F_i^{-1}(\Phi(Z_i))$ implying that X_i is distributed according to F_i . If the F_i are continuous and known, we can recreate the vectors Z_1 and Z_2 by using the pseudo-inverse CDFs on the original data ($Z_1 = \Phi^{-1}(F_1(X_1))$ and $Z_2 = \Phi^{-1}(F_2(X_2))$). We could then generate a good estimate of R using the transformed vectors Z_1 and Z_2 and maximum likelihood estimation, for example $\sum_{i=1}^N \frac{Z_1^{(i)} Z_2^{(i)}}{N}$ would be a natural estimate for the correlation. However, when either vector is not continuous, the simple pseudo-inverse transformation does not allow for correct estimation of the correlation. Now assume X_2 is a binary or ordinal variable as in many of our cases but that the marginal is not known. Instead of using a plug in value for Z_2 (say, by estimating the marginal F_2), we contend that the ranks of our latent continuous Z_2 are the same as those of the observed variable X_2 . The estimation procedure then iterates the following two steps: Using the ranks of X_1 and X_2 and the current estimate of the multivariate correlation R , we can draw values of the latent variables (Z_1 and Z_2) that preserve the rank ordering of the observed data. The second step uses the sampled underlying latent variables to sample the correlation

⁸Further details of the algorithm for estimation are available in Hoff (2007).

R . These steps are iterated until stationarity is reached. Relying on the ranks and latent scale allows us to not specify the marginal distributions of the individual variables and still arrive at a proper solution to estimating R .

When values of x_{ij} are missing at random, imputation can be performed first on the latent z_{ij} scale (since the latent variables are normal, sampling from the conditional distribution of the missing data given the observed data requires a multivariate normal draw) and are then transformed to the observed scale using the empirical cumulative density functions. As this is a Bayesian procedure we produce a 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, sbgcop, and MICE

In this section, we compare the working properties of the copula based imputation with those of `Amelia II` and `MICE` packages. We evaluate each method based on an extensive simulation study as well as an empirical example 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. Since correct estimation of the coefficients and standard errors is critical for obtaining valid statistical inference, any analysis of MI procedures 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.

We therefore use the following approach to assess the validity of an MI procedure through simulation:

1. Define a full data quantity of interest, θ . In our setting, θ is 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).

We 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 $\mathbf{X} = X_1, \dots, X_J$ some of which can be costly to measure. As such, it is common that while the outcome Y is measured for all variables, some entries of the design matrix \mathbf{X} are missing. In this simulation, we exclusively focus on this situation and restrict the missingness to the explanatory variables. We will further assume that the missingness mechanism does not allow for the missingness to depend on the outcome Y .

In this situation complete case analysis (or listwise deletion) provides an unbiased estimate of the regression coefficients; however, the reduced sample size often leads to losses in efficiency, through higher standard errors. Another disadvantage of using complete case analysis whenever the number of explanatory variables J is of moderate size is that 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 significant reduction in the standard errors; however, this can induce a slight bias. White and Carlin (2010) show through an extensive 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 to demonstrate the versatility of the copula approach

without specifying any of the marginal distributions. This is precisely the scenario we described above; the probability of enough complete cases existing to estimate the regression coefficients is effectively 0.⁹

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 $A = F_{\text{Pois}, \lambda}^{-1}(F_{\mathcal{N}}(Z_1))$ and $B = F_{\text{Pois}, \lambda}^{-1}(F_{\mathcal{N}}(Z_2))$. The data generating process thus leads to the following marginal distributions for the entries in \mathbf{X} : for $j = 1, \dots, 10$

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

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

⁹The reason is that with a high probability of missingness for each variable and a large enough number of variables, the probability of observing all variables for one particular case quickly becomes very small. Specifically, with probability of missingness p and k covariates, the probability of all observations being present for one case is $(1 - p)^k$.

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.

and another one that generates **MNAR** data sets, see Appendices A and B for details. The MI procedures we considered are only valid under the MAR assumption; however, it is useful to check how each method performs when this assumption is violated - as is often the case in practice.

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). For **Amelia II**, we used the appropriate variable transformation in accordance with the package help files. For the copula approach, we did not need to specify any distributions/transformations. Using each of the three procedures, we created 20 completed data sets that were used to estimate the regression coefficients and a corresponding 95% confidence interval.¹⁰ None of the simulations had

¹⁰Throughout the simulation, the **Amelia II** software crashed numerous times, as detailed in Table D.1 in Appendix D. Due to this the results for **Amelia II** are only on a subset of the 36,000 simulations

enough complete cases to estimate the regression coefficients using listwise deletion.

The most significant source of variation in the simulation was due to the different classes of variables, followed by the correlation and the missingness coefficient. There is only a small difference in the results obtained from the **MAR** and **MNAR** data; therefore, our discussion will focus on the former, with the figures for the latter included in Appendix C. Figures 3 and 2 illustrate how the bias, coverage and interval length, vary across the interaction of the different variable classes, the correlation, and the missingness coefficient, respectively. Overall the copula method achieved an empirical coverage rate of 93% which was much higher than that of **MICE**, 87%, and **Amelia II**, 83%. Less adversarial regimes were previously studied in White and Carlin (2010), by reducing the number of covariates in our simulation we can recover similar coverage rates for the MI procedures as are reported there. Both the copula and **MICE** methods had an absolute average bias of 0.17. **Amelia II** performed worse and had a bias of 0.25. On average, all three methods had approximately the same interval length.

The copula imputations were obtained using 10,000 iterations from Hoff's (2010) package whose convergence was checked on a subset of simulations. The lag-10 autocorrelation for the thinned chains was 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). Since the copula method is sampling from the posterior distribution which requires the MCMC algorithm to converge to the stationary distribution, its computation time depends on the rate of convergence as well as the desired number of imputations. Running multiple MCMC chains in parallel to generate independent impu-

tations can reduce the computation time. This approach is slightly slower than **Amelia II** but is substantially faster than the standard MICE algorithm where all $J - 1$ variables are used to impute the j^{th} variable. Fortunately, the copula algorithm scales well as the sample size and the number of explanatory variables increases. The copula method had the lowest bias, highest coverage rate and often the longest interval length. It is noteworthy that 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, users need to check that the model parameters fully explore the parameter space. Unlike the Bayesian copula method, there are no explicit convergence criteria that can be tracked. 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, meaning the estimated standard errors were too small. MICE also had the smallest 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 II had the lowest coverage and highest bias both on average and in most scenarios that we considered. It had the smallest average interval length of 1.23, which shows that it was systematically underestimating the variance: leading to the low coverage rates.

Figure 2 shows that the average bias and the interval length increases as a function of the

proportion of missing values. This leads to a decrease in the empirical coverage as the bias increases at a faster rate than the interval length. One notable exception was the correct coverage of the copula approach for the regression parameters of the ordinal and binomial variables, both `Amelia II` and `MICE` undercovered. Given that these types of variables are frequently encountered in social science applications, these results especially suggest that using a copula approach can lead to better statistical conclusions. Moreover, the overall simulation results indicate that when a normal distribution does not well approximate the data, then the copula approach will consistently outperform both `Amelia II` and `MICE`.

Somewhat surprisingly, there seems to be less variation in the bias and the interval length as a function of the correlation, as is shown in Figure 3. Except for the normally distributed variables, the bias decreases as the correlation increases due to the reduction in the relative loss of information from the missing data.

Breaking the **MAR** assumption did not lead to drastically worse results. We observe 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.

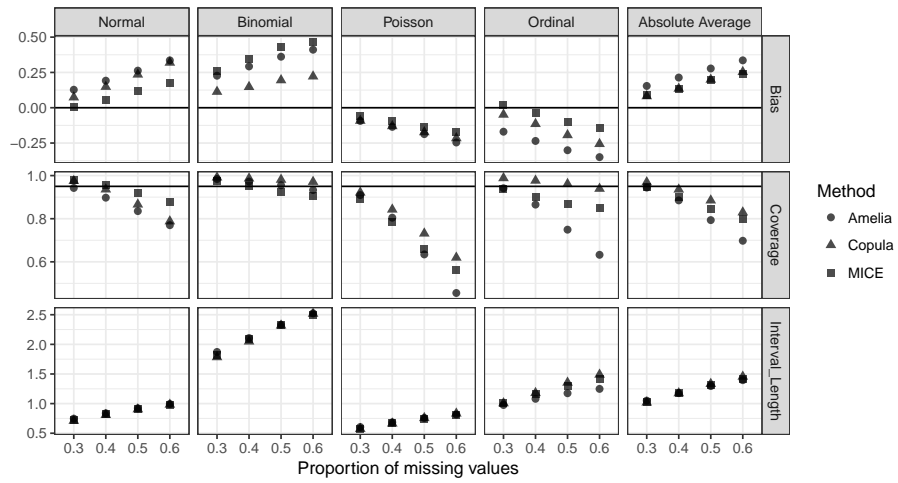


Figure 2: Simulation study results for the **MAR** data as a function of the missingness coefficient, averaging over the correlation. The plot is split by the different variable types (normal, binomial, Poisson and ordinal) and the three outcomes of interest (bias, coverage and interval length). The rightmost panel shows the result averaging over the different variable types.

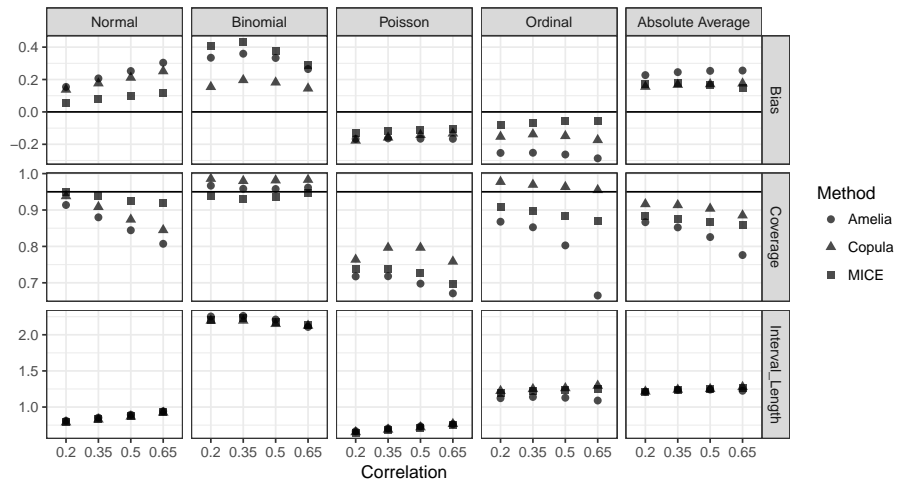


Figure 3: Simulation study results for the **MAR** data as a function of the correlation, averaging over the missingness coefficient. The plot is split by the different variable types (normal, binomial, Poisson and ordinal) and the three outcomes of interest (bias, coverage and interval length). The rightmost panel shows the result averaging over the different variable types.

5 Application Study

In this section, we provide a comparison of the three imputation methods using an application from political science. The empirical example shows how copula methods can be used to generate imputations in 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 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. We can see that many of the variables have a large share of missing observations. If instead of multiple imputations, the authors used in listwise deletion then the number of observation in the regression model would have been approximately halved. Instead, Krieckhaus et al. (2014) use *Amelia II* 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 the data set includes different types of variables, continuous, binary, as well as ordinal. We created 20 multiple imputed data sets using each of the imputation techniques: *Amelia II*, *MICE*, and *sbgcop*. 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. 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. Note that, again, we do not have to declare any of the variable types or make any other specification or transformation of the data.

Figure 4 shows the coefficient estimates and 95% intervals for the replicated model based on each of the imputation techniques, as well as when list-wise deletion is used. First, the results are quite similar for the *Inequality*, *Income*, and *Age* variables. Even for the models based on listwise deletion. 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 for the other variables included in the model. 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.

Based on the data imputed with the copula method, the estimated association of *Institutional Confidence* with *Democracy Satisfaction* is significantly stronger compared to the models based on listwise deletion or other imputation methods. Similarly, the estimated effect of *Leftist Ideology* is also substantially larger according to the copula imputed data. On the other hand, the association of *Education* levels with *Democracy Satisfaction* is significantly smaller. Based on the copula, the relationships of *Interest in Politics*, and *Prior Regime Evaluation* with the dependent variable of *Democracy Satisfaction* are all modeled to be weaker, compared to the other methods (and the non-imputed data), though the confidence intervals overlap.

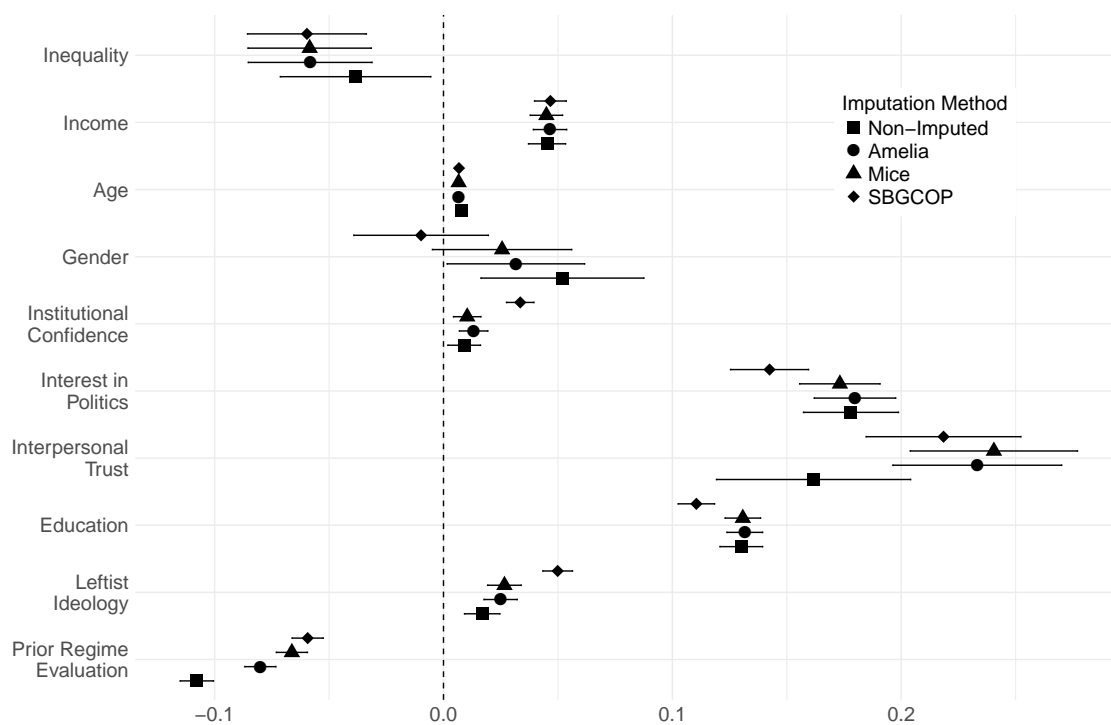


Figure 4: Coefficient estimates and confidence intervals for *Model 1* in *Table 1* in Krieckhaus et al. (2014) based on three imputation techniques and list-wise deletion

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.

Based on the simulation results, especially with respect to binary and ordinal variables, and the theoretical properties we are confident in the accuracy of the copula imputation method. These results suggest then that *Gender* is not associated with people's satisfaction with democracy, whereas *Institutional Confidence* and *Left* ideology both have much stronger effects.

6 Conclusion

What practical lessons can we learn about how to deal with missing data? 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 little pre-specification of the data. With the rank based approach introduced by Hoff (2007), the Gaussian copula can be used to impute binary, ordinal, and continuous variables. We discuss the theoretical properties of the copula method and its theoretical attractiveness compared to other commonly employed techniques. In particular, the Gaussian copula introduced here enables researchers to make imputation via draws from a valid posterior of the joint distribution without specifying the distributions of the individual variables. Moreover, we present evidence from simulations that it performs better than either `Amelia II` or `MICE`, especially when it comes to non-normally distributed data.

While the three imputation methods perform relatively similarly, throughout the simulation, the Copula method does have the lowest average bias (tied with `MICE`) and the highest coverage rate (93%). More so, `MICE` requires specification of the conditional distributions whereas the copula method does not. Recent theoretical results for `MICE` suggest that good performance heavily relies on being approximately correct in the choice of conditionals (Li et al., 2012). On the other hand, 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 aspect that we have not addressed herein is the validity and sensitivity to the unassessable assumptions made when analyzing data with missing values (Molenberghs et al., 2014), *i.e.* the type of missingness mechanism. Bojinov et al. (2017) show that the Gaussian copula approach can be used to assess the validity of the missing always at random assumption (a slightly stronger assumption that implies **MAR**). Their results suggest that by using a Gaussian copula for generating imputations, the analyst can also easily diagnose the assumptions they made and quickly identify variables which are likely to break these assumptions. This adds another benefit to using the method discussed in this paper.

Consideration must also be given to the computational cost of any procedure. As indicated by Graham (2009) the disadvantages of EM approaches are especially large when imputing databases with many variables or applications of “big data”. MICE can be computationally less expensive but suffers when the number of variables increases as the correct choice for each of the conditionals becomes increasingly unlikely. The semiparametric copula approach described here relies on MCMC, its speed does not depend on the fraction of missing data and scales nicely in the dimension of the dataset. This makes it possible to impute even large database in a relatively timely manner and no pre-specification of the data. Moreover, using the copula model to multiply 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. 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.

SUPPLEMENTARY MATERIAL

Code will be provided on the author's Dataverse.

R-packages for Imputation: 3 R-packages used to impute the missing data: Amelia II, MICE, sbgeop

R-code for simulation in Section 4: R-code to replicate simulation study in section 4.

R-code for Application in Section 5: R-code to replicate application in section 5.

A Missing at Random

We now describe a missing data mechanism that always produces **MAR** data. Our goal is to make the simulations as realistic as possible; therefore some variables will be fully observed, and others will have different amounts of missing values.

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 $\mathbf{X}_{cc}^{(1)}$ be the complete cases after removing the any rows that have missing values.
3. The probability of the i^{th} observation missing for the remaining thirty two variables is proportional to a logistic regression on the fully observed $\mathbf{X}_{cc}^{(1)}$. The probabilities are then 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 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.

The proportion of missing values is slightly lower than the MC as four variables are fully observed, and four others only have 5-6% of their values missing.

B Missing not at Random

We now describe a missing data mechanism that produces **MNAR** data with extremely high probability.

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

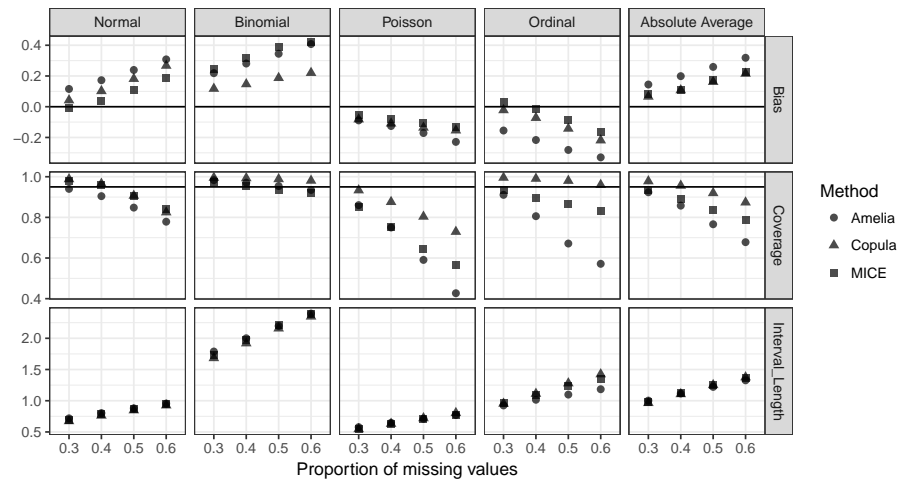


Figure C.1: Simulation study results for the **MNAR** data as a function of the missingness coefficient, averaging over the correlation. The plot is split by the different variable types (normal, binomial, Poisson and ordinal) and the three outcomes of interested (the bias, coverage and interval length). The rightmost panel shows the result averaging over the different variable types.

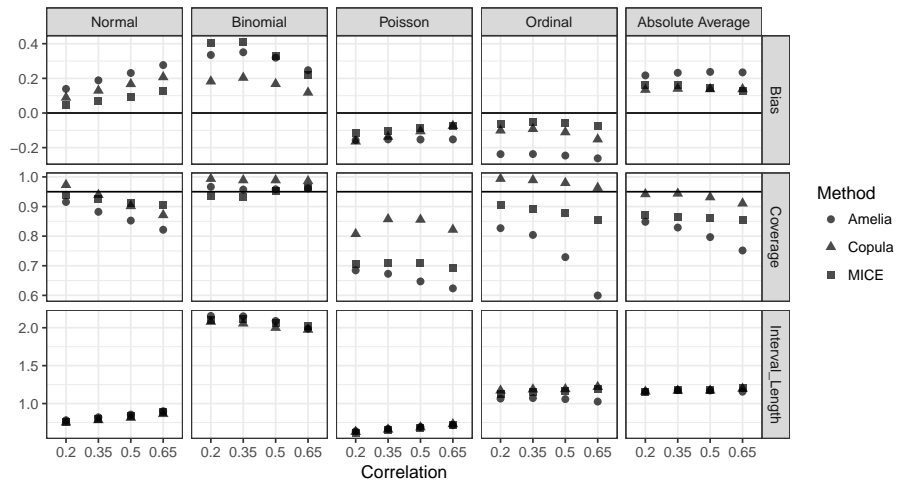


Figure C.2: Simulation study results for the **MNAR** data as a function of the correlation, averaging over the missingness coefficient. The plot is split by the different variable types (normal, binomial, Poisson and ordinal) and the three outcomes of interested (the bias, coverage and interval length). The rightmost panel shows the result averaging over the different variable types.

D Number of Simulations for which Amelia II 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 II* crashes out of the 1000 simulations under each of the possible scenarios.

E Example *sbgcop* Application

In this section, we discuss how to use the ‘*sbgcop*’ package for multiple imputation in the context of conducting inferential analysis on data with missingness. Specifically, we show how to conduct regression analysis in the presence of missing data using an example dataset. First we simulate a dataset in which we introduce missingness.

```
1  # simulate data
2  set.seed(6886)
3  n <- 100
4  x1 <- rnorm(n) ; x2 <- rnorm(n) ; x3 <- rnorm(n)
5  y <- 1 + 2*x1 -1*x2 + 1*x3 + rnorm(n)
6
7  ## organize into matrix
8  raw <- cbind(y, x1, x2, x3)
9
10 ## simulate missingness
```

```
11 naMat <- matrix(rbinom(n*4,1,.7),
12     nrow=nrow(raw),ncol=ncol(raw))
13 naMat[naMat==0] <- NA
14
15 ## remove observations
16 data <- raw * naMat
17
18 ## summarize missingness
19 missStats <- apply(data, 2, function(x){sum(is.na(x))/nrow(data)})
20 missStats <- matrix(missStats,
21     ncol=1,
22     dimnames=list(colnames(data),'Prop. Missing')
23     )
```

Using this simulated dataset, our goal is to show how to conduct inference on the effect of x_1 , x_2 , and x_3 on y after imputing the missing values with the `sbgcop` package in R. `sbgcop` is available on CRAN and can be installed and loaded into your R session just as any other package.

```
24 install.packages('sbgcop')
25 library(sbgcop)
```

The key function in this package is `sbgcop.mcmc` and there are four arguments that should always be set (for a full list of arguments run `?sbgcop.mcmc`):

- `Y`: a matrix with missing values to be imputed
- `nsamp`: number of iterations of the Markov chain
- `odens`: number of iterations between saved samples
- `seed`: an integer for the random seed

The Y argument specifies the dataset to be imputed. The object passed to the argument must be in **matrix** format. Additionally, users should only include variables that can provide information to the imputation algorithm. For example, this can include lags and leads of a variable in the case of time-series-cross-sectional data. Identification variables, such as actor names, abbreviations, or years, should not be included in the **matrix**.

The imputation procedure in `sbgcop.mcmc` is a Bayesian estimation scheme, so users must pass the number of iterations for which they want the Markov chain to be run to the `nsamp` argument. If `nsamp` is set to 100, then the Markov chain will run for 100 iterations and 100 imputed datasets will be produced. The `odens` argument specifies how often an iteration from the Markov chain should be saved. Thus, if `nsamp` is set to 100 and `odens` is set to 4, 25 imputed datasets will be returned by `sbgcop.mcmc`. Last, since this is a Bayesian model and we will be sampling from distributions to arrive at parameter values, one should always pass an integer to the `seed` argument. This way when users rerun `sbgcop.mcmc` they will arrive at the same results.

To impute missingness in our example dataset, we pass our `data` object to the `sbgcop.mcmc` function. We run the Markov chain for 2000 iterations and save every 10th iteration. We store the output from `sbgcop.mcmc` to `sbgcopOutput`.

26

```
sbgcopOutput <- sbgcop.mcmc(Y=data, nsamp=2000, odens=10, seed=6886)
```

This is quite simple to do as the output from `sbgcop.mcmc` is simply a list. The first element in this list is `C.psamp`, which contains posterior samples of the correlation matrix. The `C.psamp` is structured as an array of size $p \times p \times \text{nsamp}/\text{odens}$. Where p indicates the number of variables included in the imputation process. In our case, the `data` object

includes 4 variables and we ran the Markov chain for 2000 iterations saving every tenth. Thus giving us dimensions of: 4 x 4 x 200.

Each value in this array is providing us with the estimated association between a pair of parameters at every saved iteration of the Markov chain. We show an example below using the 100th and 200th saved iterations.

```

27 sbgcopOutput$C.psamp[, ,c(100,200)]
28
29 ## , , 100
30 ##
31 ##           y           x1           x2           x3
32 ## y  1.0000000  0.78961179 -0.43494151  0.36593885
33 ## x1  0.7896118  1.00000000 -0.08686933  0.05172101
34 ## x2 -0.4349415 -0.08686933  1.00000000 -0.14619182
35 ## x3  0.3659389  0.05172101 -0.14619182  1.00000000
36 ##
37 ## , , 200
38 ##
39 ##           y           x1           x2           x3
40 ## y  1.0000000  0.68269537 -0.46139236  0.4138161
41 ## x1  0.6826954  1.00000000  0.08754115  0.1495993
42 ## x2 -0.4613924  0.08754115  1.00000000 -0.1278238
43 ## x3  0.4138161  0.14959933 -0.12782384  1.0000000

```

To generate a trace plot of this data we need to restructure our dataframe into a long format. We can do so using the `reshape2` package:

```

44 library(reshape2)
45 sbgcopCorr = reshape2::melt(sbgcopOutput$'C.psamp')
46
47 # remove cases where variable is the same in both columns
48 sbgcopCorr = sbgcopCorr[sbgcopCorr$Var1 != sbgcopCorr$Var2,]
49
50 # construct an indicator for pairs of variables
51 sbgcopCorr$v12 = paste(sbgcopCorr$Var1, sbgcopCorr$Var2, sep='-')
52
53 #
54 print(head(sbgcopCorr))
55
56 ##   Var1 Var2 Var3      value  v12
57 ## 2   x1   y     1  0.62439270 x1-y
58 ## 3   x2   y     1 -0.43347850 x2-y
59 ## 4   x3   y     1  0.28013565 x3-y
60 ## 5    y   x1     1  0.62439270 y-x1
61 ## 7   x2  x1     1  0.03581958 x2-x1
62 ## 8   x3  x1     1  0.15626246 x3-x1

```

Using the `reshape2` package we have reformatted the array into a dataframe, in which the first two columns designate the variables for which a correlation is being estimated, the third an indicator of the saved iteration, the fourth the correlation, and the fifth an indicator designating the variables being compared.

Next, we use `ggplot2` to construct a simple trace plot shown in Figure E.3.

```
63 library(ggplot2)
64
65 ggplot(sbgcopCorr, aes(x=Var3, y=value, color=v12)) +
66   geom_line() +
67   ylab('Correlation') + xlab('Iteration') +
68   facet_wrap(~v12) +
69   theme(legend.position='none')
```

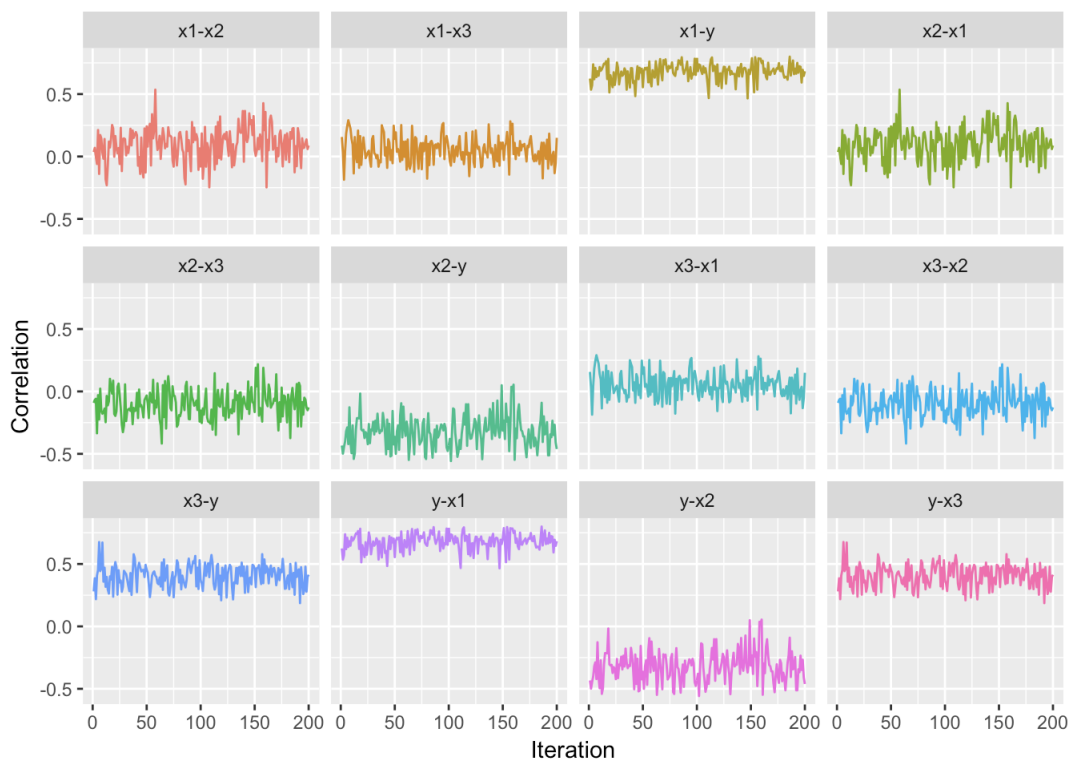


Figure E.3: Trace plot of correlation between variables.

Based on these trace plots we can see that the Markov chain tends to converge rather quickly in this example. The `coda` package provides an excellent set of diagnostics to test convergence in more depth.

After conducting the imputation and evaluating convergence, our goal is now to use the imputed datasets to conduct inferential analysis. For the purpose of this example, we estimate the effect of x_1 , x_2 , and x_3 on y . By using `sbgcop` as above we have generated 200 copies of our original dataset in which posterior samples of the original missing values have been included. Each of these copies are saved in the output from `sbgcop.mcmc`, which has dimensions of 100 x 4 x 200.

The first two dimensions of this object correspond to the original dimensions of our `data` object, and the third corresponds to the number of saved iterations from the Markov chain.

Having generated a set of imputed datasets, our next step is to use a regression model to estimate the effect of our independent variables on y . We cannot just use one of the imputed datasets – as this would not take into account the uncertainty in our imputations. Instead we run several regression on as many of the imputed datasets generated by `sbgcop.mcmc` that we think are appropriate. For the sake of this example, we utilize all 200 imputed datasets, but typically randomly sampling around 20 imputed datasets should be sufficient.

Each time we run the regression model, we will save the coefficient and standard errors for the independent variables and organize the results into a matrix as shown below.

```
70 coefEstimates <- NULL
71 serrorEstimates <- NULL
```

```

72 for( copy in 1:dim(sbgcopOutput$'Y.impute')[3]){
73     # extract copy from sbgcopOutput
74     copyDf <- data.frame(sbgcopOutput$'Y.impute'[, ,copy])
75     names(copyDf) <- colnames(sbgcopOutput$Y.pmean)
76     # run model
77     model <- lm(y~x1+x2+x3,data=copyDf)
78     # extract coefficients
79     beta <- coef(model)
80     coefEstimates <- rbind(coefEstimates, beta)
81     # extract standard errors
82     serror <- sqrt(diag(vcov(model)))
83     serrorEstimates <- rbind(serrorEstimates, serror)
84 }
85
86 print(head(coefEstimates))
87
88 ##      (Intercept)      x1      x2      x3
89 ## beta  0.6576411  1.449662 -1.1290934  0.4569379
90 ## beta  0.7436243  1.661250 -1.0542155  0.6866980
91 ## beta  0.8299671  1.613892 -1.1363969  0.7454211
92 ## beta  0.8073597  1.513452 -0.7512275  0.6331863
93 ## beta  0.8112010  1.583065 -0.9608251  0.6529509
94 ## beta  0.7882072  1.509635 -0.5152139  0.8897130

```

The last step is to combine each of the estimates using using Rubin's rule. Many existing packages have implemented functions to aid in this last step, one could use the pool function from mice or the mi.meld function from Amelia II as below.

```

95 paramEstimates <- Amelia::mi.meld(q=coefEstimates, se=serrorEstimates)
96 print(paramEstimates)
97
98 ## lq.mi
99 ##      (Intercept)      x1      x2      x3
100 ## [1,]  0.892732  1.70032 -0.9023761  0.7235922
101
102 ## lse.mi
103 ##      (Intercept)      x1      x2      x3
104 ## [1,]  0.1680402  0.1965969  0.2213771  0.1588638

```

The resulting parameter estimates take into account the uncertainty introduced through

the imputation process, and we can interpret them just as we would interpret the results from a typical regression.

Below we show the full set of steps required to conduct a regression analysis in the context of missing data using `sbgcop`.

```
1 library(sbgcop)
2 sbgcopOutput <- sbgcop.mcmc(Y=data, nsamp=2000, odens=10, seed=6886)
3
4 ## restructure posterior samples of correlation matrix
5 library(reshape2)
6 sbgcopCorr = reshape2::melt(sbgcopOutput$'C.psamp')
7 sbgcopCorr = sbgcopCorr[sbgcopCorr$Var1 != sbgcopCorr$Var2,]
8 sbgcopCorr$v12 = paste(sbgcopCorr$Var1, sbgcopCorr$Var2, sep='-')
9
10 ## trace plot of C.psamp
11 library(ggplot2)
12 ggplot(sbgcopCorr, aes(x=Var3, y=value, color=v12)) +
13   geom_line() +
14   ylab('Correlation') + xlab('Iteration') +
15   facet_wrap(~v12) +
16   theme(legend.position='none')
17
18 ## conduct regression analysis
19 coefEstimates <- NULL
20 serrorEstimates <- NULL
21 for( copy in 1:dim(sbgcopOutput$'Y.impute')[3]){
22   copyDf <- data.frame(sbgcopOutput$'Y.impute'[, ,copy])
23   names(copyDf) <- colnames(sbgcopOutput$Y.pmean)
24   model <- lm(y~x1+x2+x3,data=copyDf)
25   beta <- coef(model)
26   coefEstimates <- rbind(coefEstimates, beta)
27   serror <- sqrt(diag(vcov(model)))
28   serrorEstimates <- rbind(serrorEstimates, serror) }
29
30 ## combine estimates using Rubin's rules
31 paramEstimates <- Amelia::mi.meld(q=coefEstimates, se=serrorEstimates)
```

References

- Bojinov, I., N. Pillai, and D. Rubin (2017). Diagnosing missing always at random in multivariate data. *arXiv preprint arXiv:1710.06891*.
- Casella, G. and E. I. George (1992). Explaining the Gibbs Sampler. *The American Statistician* 46(3), 167–174.
- Chen, S.-H. and E. H. Ip (2015). Behaviour of the Gibbs sampler when conditional distributions are potentially incompatible. *Journal of statistical computation and simulation* 85(16), 3266–3275.
- 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.
- 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. Lesafre (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 (2012). AMELIA II: A Program for Missing Data – Documentation.
- 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. (2006). Imputation algorithm using copulas. *Metodoloski zvezki* 3(1), 109.
- 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 Political 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.
- Kriekhaus, 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.
- Marini, M. M., A. R. Olsen, and D. B. Rubin (1980). Maximum-likelihood estimation in panel studies with missing data. *Sociological methodology* 11, 314–357.
- 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.
- R Development Core Team (2004). *R: A language and environment for statistical computing*. Vienna, Austria.
- Robbins, M. W., S. K. Ghosh, and J. D. Habiger (2013). Imputation in High-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey. *Journal of the American Statistical Association* 108(501), 81–95.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* 63(3), 581–592.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. 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.
- 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 (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software 45*(3), 1–67.
- Van Buuren, S. and K. Groothuis-Oudshoorn (2011). MICE: Multivariate imputation by chained equations in R. *Journal of statistical software 45*(3), 1–67.
- 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.