OUTLINE OF PAPER 1
High Dimensional Imputation for the Social Sciences: A Comparison of State-of-the-Art Methods

This is the outline planned for the paper. It will not be included in the final work.

1. **Introduction:** Frame problem; Discuss background literature; Focus/Reason to write paper; Content Summary.
2. **Algorithms and Imputation methods:** Describe bridge, blasso, DURR, IURR, MI-PCA, etc.; Focus on minimal possible description to give reader sense of what the method is (max Deng et al. (2016); Reference papers for details.
3. **Simulation Studies**
    - Simulation Study Procedure
        - Step 1: Data Generation
        - Step 2: Missing Data Imposition
        - Step 3: Imputation
        - Step 4: Analysis
    - Conditions
    - Comparison Criteria
    - Results: distinguish by type of performance measure
        - Experiment 1
        - Experiment 2
4. **Resampling Study**
    - Resampling Study Procedure
        - Data preparation: documentation for the data; what is it; why collected; general original demographics of cases; selected demographics (e.g. western European Countries); systematic cleaning process with general purpose; reference to appendix for details.
        - Analysis models
        - Missing data imposition
    - Results: again divide by type.
        - Bias
        - Confidence Interval Coverage
        - Imputation Time
5. **Discussion:** Synthesize findings, make parallels and comparisons.
6. **Conclusions:** Short take home message, limitations, future directions (hint at MY future work)
7. Appendices
    - Extra Results
    - Methods Details
    - EVS quirks

High Dimensional Imputation for the Social Sciences
A Comparison of State-of-the-Art Methods

Edoardo Costantini

January 22, 2021

1 Introduction

Today's social, behavioral and medical scientists are blessed with a wealth of large, high-quality data that can help investigate the complex relationships between social, psychological and biological factors in shaping individual and societal outcomes. Large social scientific datasets, such as the World Values Survey (WVS), or the European Values Study (EVS), are easily available and initiatives have been undertaken to link and extend these datasets into a full systems of linked open data (LOD).

Making use of the full potential of these data sets requires dealing with the crucial problem of multivariate missing data. Rubin's Multiple Imputation approach (Rubin, 1987) was developed to specifically address the issue of missing responses in surveys. The basic idea underlying MI is to repeatedly sample replacement values for each missing data point, by sampling from a predictive distribution given observed data. This procedure leads to the definition of multiple datasets, each imputed with different samples from the predictive distribution, that can be analyzed separately using standard complete-data analyses. Results can then be pooled following Rubin's rules (Rubin, 1987).

Multiple Imputation relies on the crucial Missing At Random (MAR) assumption. Meeting this assumption requires specifying imputation models for the MI procedure that include all observed variables that are correlates of missingness. Omitting an observed predictor related to missingness from an imputation model might lead to substantial bias in the estimation of any analysis model, and invalidate hypothesis testing involving the imputed variables.

As a result, when it comes to defining the set of auxiliary variables for MI imputation models, an inclusive strategy (i.e. including numerous auxiliary variables) is generally preferred: compared to restrictive approach (i.e. including few or no auxiliary variables), it reduces the chances of omitting important correlates of missingness, making the MAR assumption more plausible. Furthermore, the inclusive strategy has been shown to reduce estimation bias and increase efficiency (Collins et al., 2001), as well as reducing the chances of specifying uncongenial imputation and analysis models (Meng, 1994).

In practice, however, an inclusive strategy increases the dimensionality of the imputation models and identification and computational limitations often force researchers to make arbitrary decisions on what predictors to include in the imputation models. One serious risk of an inclusive strategy is the occurrence of singular matrices within the imputation algorithm. When data is high-dimensional (n is *not* substantially larger than p) or afflicted by high collinearity (correlation among certain variables is so high that some of their linear combinations have no variance) the data covariance matrix is singular. Singular matrices are not invertible, an operation that is fundamental in the estimation of the imputation models in any parametric Multiple Imputation procedure. As a result, the possible high dimensionality of the observed data matrix, resulting from an inclusive strategy, can prevent a straightforward application of MI algorithms, such as MICE (van Buuren, 2012), or force researcher to make difficult choices regarding which variables to use.

(Background) Recent developments in high dimensional multiple imputation techniques represent interesting opportunities to embrace an inclusive strategy without facing the risk of having too many superfluous auxiliary variables in the imputation models. Some researchers have focused on high-dimensional single imputation methods in an effort to improve the accuracy of individual imputations (Kim et al., 2005; Stekhoven and Bühlmann, 2011; D'Ambrosio et al., 2012). However, the main task of social scientists is to make inference about a population based on a sample of observed data points, and single imputation is simply inadequate for this purpose: it does not guarantee unbiased and confidence valid estimates of the parameters of interest (Rubin, 1996).

Multiple Imputation is more suitable for the type of research social scientists are involved in. Its combination with high dimensional prediction models has been directly tackled by specific algorithms combining MICE with shrinkage methods (Zhao and Long, 2016; Deng et al., 2016), dimensionality reduction methods (Song and Belin, 2004; Howard et al., 2015), and even non-parametric prediction trees (Reiter, 2005; Burgette and Reiter, 2010; Doove et al., 2014; Shah et al., 2014).

Although some of these approaches have been tested in proper high-dimensional contexts, many of them have been either proposed or tested exclusively for low-dimensional imputation settings. These methods have the potential of simplifying the decisions a social scientist needs to make when dealing with missing values, but so far there has been little research on their comparative performances when applied to data social scientists actually want to use.

Scope With this article, we set out to provide a comparison of state-of-the-art high-dimensional imputation algorithms that do not require the researcher to make decisions on which variables to include in the procedure. We compared imputation methods based on their ability to allow inferential statements that are as statistically valid as if they were made on a dataset without missing data. Hence, in assessing the methods performances, the primary focus of this article was the *statistical validity* (Rubin, 1996) of the substantive analysis performed on data treated with different high-dimensional MI procedures. The comparison was developed through two simulation studies and a resampling study using real survey data.

Outline This paper is organized as follows. Section 2 discusses the imputation methods compared. Section 3 presents the two simulation studies, their design and the result of the comparison. Section 4 presents the resampling study performed on the 2017 wave of the EVS. Section 5 discusses the implication of the combined results of the simulation and resampling studies. Finally, section 6 provides concluding remarks, a description of the limitations of the study, and future research directions we want to take.

2 Imputation methods and Algorithms

Consider a dataset $\boldsymbol{Z}$ of dimensionality $n \times p$, with n observations (rows) and p variables (columns). Assume there are t ($t < p$) variables with missing values in $\boldsymbol{Z}$ and these t variables are part of some substantive model of scientific interest (e.g. some linear regression model). An imputation procedure targeting these t variables could be used to allow fitting the substantive model without discarding data units (rows).

Let z_j denote one of these t variables, and let $z_{j,obs}$ and $z_{j,mis}$ denote its observed and missing components. Let $\boldsymbol{Z}_{-j}$ be the collection of $p-1$ variables in $\boldsymbol{Z}$ excluding z_j, and denote $\boldsymbol{Z}_{-j,obs}$ and $\boldsymbol{Z}_{-j,mis}$ the components of $\boldsymbol{Z}_{-j}$ corresponding to the data units in $z_{j,obs}$ and $z_{j,mis}$, respectively.

$\boldsymbol{Z}_{-j}$ contains a $p-t$ subset of variables that are not target of imputation and constitute a pool of possible *auxiliary* variables that could be used to improve the imputation procedure. Let $\boldsymbol{A}$ denote this set of auxiliary variables.

2.1 Multiple Imputation by Chained Equations

The general Multiple Imputation framework considered here is the Fully Conditional Specification of imputation models as implemented in the Multiple Imputation by Chained Equations algorithm proposed by van Buuren and Groothuis-Oudshoorn (2011).

Assume that $\boldsymbol{Z}$ is the result of n random samples from a multivariate distribution defined by an unknown set of parameters $\boldsymbol{\theta}$. The chained equations approach proposes to impute values by iteratively sampling from the conditional distributions of the t variables with missing values $P(z_1|\boldsymbol{Z}_{-1}, \boldsymbol{\theta}_1)...P(z_t|\boldsymbol{Z}_{-t}, \boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}_1...\boldsymbol{\theta}_t$ are imputation model parameters specific to the conditional densities of each variable with missing values.

More precisely, the MICE algorithm takes the form of a Gibbs sampler where the m-th iteration $m = (1, ..., M)$ successively draws, for each j-th target variable ($j = 1, ..., t$), from the following distributions:

$$\hat{\boldsymbol{\theta}}_j^{(m)} \sim p(\boldsymbol{\theta}_j|\boldsymbol{z}_{j,obs}, \boldsymbol{Z}_{-j,obs}^{(m)}) \tag{1}$$

$$z_{j,mis}^{(m)} \sim p(z_{j,mis}|\boldsymbol{Z}_{j,mis}^{(m)}, \hat{\boldsymbol{\theta}}_j^{(m)}) \tag{2}$$

$\hat{\boldsymbol{\theta}}_j^{(m)}$ and $z_{j,mis}^{(m)}$ are draws from the parameters posterior distribution (1) and the missing data posterior predictive distribution (2), respectively. After convergence, D different sets of imputations are kept to form D differently imputed data sets. Any substantive model can then be fit to each data, and estimates can be pooled appropriately.

The methods described below follow these general framework but differ in the building blocks they use to define the distributions in (1) and (2).

2.1.1 MICE with Bayesian Ridge (bridge)

The *bridge* imputation procedure uses as building block the Bayesian imputation under the normal linear model, with standard non-informative priors for each parameter, as presented by (van Buuren, 2012, p. 68, algorithm 3.1).

In this approach, the sampling of each $\hat{\boldsymbol{\theta}}_j^{(m)}$ in (1) relies on the inversion of the cross-product of the observed data matrix $\boldsymbol{Z}_{j,mis}^{(m)}$. By adding a biasing ridge penalty κ, singularity is circumvented and the sampling scheme is possible even on data that is high-dimensional or afflicted by high collinearity.

The value of κ is usually chosen close to zero (e.g. $\kappa = 0.0001$), as values larger than .1 may introduce systematic bias. However, larger values may be necessary to invert the observed data matrices in certain scenarios. In the present work, the value of κ was decided by means of a cross-validation procedure described below.

2.1.2 MICE with Bayesian lasso (blasso)

A Bayesian Lasso linear model for data analysis is a regular Bayesian multiple regression with prior specifications for the regression coefficients that induces some form of shrinkage toward 0 of the sampled parameters values (Park and Casella, 2008; Hans, 2009) effectively performing a form of Bayesian model selection.

Given data with sample size n, consider the dependent variable y and a set of predictors X, the Bayesian Lasso linear regression specification, used within the blasso imputation algorithm, is that specified by Hans (2010):

$$p(y|\beta,\sigma^2,\tau) = N(y|X\beta,\sigma^2 I_n) \tag{3}$$

$$p(\beta_j|\tau,\sigma^2,\rho) = (1-\rho)\delta_0\beta_j + \rho\left(\frac{\tau}{2\sigma}\right) \times exp\left(\frac{-\tau\left\|\beta\right\|_1}{\sigma}\right) \tag{4}$$

$$\sigma^2 \sim IG(a,b) \tag{5}$$

$$\tau \sim G(r,s) \tag{6}$$

$$\rho \sim Beta(g,h) \tag{7}$$

The expression in (3) represents the density function, of a multivariate normal random variable with mean $X\beta$ and covariance matrix $\sigma^2 I_n$, evaluated at y. The prior expressed in (4) is the expansion on Park and Casella (2008) double exponential prior developed by Hans (2010) to accommodate for uncertainty regarding, not only the value of the regression coefficients, but also the model sparsity. Finally, equations (5) to (7) represent hyper priors for the residual variance σ^2, the penalty parameter τ, and the sparsity parameter ρ, respectively.

The Bayesian Lasso imputation algorithm used here is a standard Multiple Imputation MCMC sampler that replaces (1) with the full conditional distributions computed by Hans (2010), based of the prior specifications in (5) to (7). Posterior parameters draws are then used to sample plausible values from the predictive distributions of the missing data.

The R code to perform blasso imputation is heavily based on the Bayesian Lasso R Package *blasso* developed by Hans (2010) and can be found on the author's GitHub page `https://github.com/EdoardoCostantini/imputeHD-comp`. For a detailed description of the algorithm for Bayesian Lasso Multiple Imputation (blasso) in a univariate missing data context we recommend reading Zhao and Long (2016).

2.1.3 Direct Use of Regularized Regression (DURR)

As proposed by Zhao and Long (2016) and Deng et al. (2016), Frequentist Regularized Regression can be directly used in a MICE algorithm to perform multiple imputation of high dimensional data. At iteration m, for a target variable $\boldsymbol{z}_j$, the DURR algorithm uses these building blocks within the MICE framework:

- Generate a bootstrap sample $\boldsymbol{Z}^{*(m)}$ by sampling with replacement rows of $\boldsymbol{Z}$, and train a regularized linear regression model (such as Lasso regression) with $\boldsymbol{z}_{j,obs}$ as outcome and $\boldsymbol{Z}^{*(m)}_{-j,obs}$ as predictors. This produces a set of parameter estimates (regression coefficients and error variance) $\hat{\boldsymbol{\theta}}^{(m)}_j$ that can be considered as sampled from (1).

- Predict $\boldsymbol{z}_{j,mis}$, based on $\boldsymbol{Z}^{*(m)}_{-j,mis}$ and $\hat{\boldsymbol{\theta}}^{(m)}_j$, to obtain draws from the posterior predictive distribution of the missing data (2).

2.1.4 Indirect Use of Regularized Regression (IURR)

While DURR performs simultaneously model trimming and parameter estimation in (1), another approach is to use regularized regression exclusively for model trimming, and to follow it with a standard multiple imputation procedure (Zhao and Long, 2016; Deng et al., 2016). At iteration m, the IURR algorithm performs the following steps for each target variable:

- Fit a multiple linear regression using a regularized regression method with $\boldsymbol{z}_{j,obs}$ as dependent variable and $\boldsymbol{Z}^{(m)}_{-j,obs}$ as predictors (compared to DURR, the original data is used, not a bootstrap sample). In this model, the regression coefficients that are

not shrunk to 0 identify the active set of variables that will be used as predictors in the actual imputation model.

- Obtain Maximum Likelihood Estimates of the regression parameters and error variance in the linear regression of $\boldsymbol{z}_{j,obs}$ on the active set of predictors in $\boldsymbol{Z}^{(m)}_{-j,obs}$ and draw a new value of these coefficients by sampling from a multivariate normal distribution centered around these MLEs

$$(\hat{\boldsymbol{\theta}}^{(m)}_j, \hat{\sigma}^{(m)}_j) \sim N(\hat{\boldsymbol{\theta}}^{(m)}_{MLE}, \hat{\boldsymbol{\Sigma}}^{(m)}_{MLE}) \tag{8}$$

  so that (8) corresponds to (1) in the general MICE framework.

- Impute $\boldsymbol{z}_{j,mis}$ by sampling from the posterior predictive distribution based on $\boldsymbol{Z}^{(m)}_{j,mis}$ and the parameters posterior draws $(\hat{\boldsymbol{\theta}}^{(m)}_j, \hat{\sigma}^{(m)}_j)$.

2.1.5 MICE with PCA (MI-PCA)

By extracting Principal Components from the auxiliary variables, it is possible to summarise the information contained in this set with just a few components, and then perform a standard MICE algorithm in a low dimensional setting. The MI-PCA imputation procedure can be summarized as follows:

- Extract the first principal components that cumulative explain at most 50% of the variance in the auxiliary variables $\boldsymbol{A}$, and collect them in a new data matrix $\boldsymbol{A}^{'}$;
- Create a new data matrix $\boldsymbol{Z}^{'}$ by replacing the subset of auxiliary variables $\boldsymbol{A}$ with $\boldsymbol{A}^{'}$
- Use the standard MICE algorithm with the Bayesian imputation under the normal linear model, using standard non-informative priors (van Buuren, 2012, p. 68, algorithm 3.1) as building block to obtain multiply imputed datasets from the low dimensional $\boldsymbol{Z}^{'}$.

Note that if missing values are present in the set of auxiliary variables, one can fill them in with a stochastic single imputation algorithm of choice as the goal of said imputation would be to simply allow PCs extraction and not inferential. This method is inspired by Howard et al. (2015) and the *PcAux* R-package (Lang et al., 2018) that implements and developed its ideas.

2.1.6 MICE with regression trees (MI-CART and MI-RANF)

MI-CART is a method that uses as building blocks for the MI Gibbs sampler non parametric classification and regression trees (CART). At the m-th iteration, for a target variable z_j, a CART predictive model is trained on the observed part of the data, which corresponds to equation 1 in the general MICE framework. This results in a tree with several leaves, each containing a subset of $z_{j,obs}$. All units with a missing value in z_j are then put down this tree and end up in one of the leaves. Finally, one value is randomly selected from the subset of $z_{j,obs}$ in this leaf and used for imputation, a step that corresponds to 2 in the general MICE framework.

The implementation of MI-CART used in this paper corresponds to the one presented in (Doove et al., 2014, p. 95, algorithm 1) and the *impute.mice.cart()* R function from the mice package.

The Multiple Imputation with Random Forest algorithm (MI-RANF) is only slightly different from MI-CART, with the main difference residing in how the donors pool is defined. In MI-RANF, the building block for equation 1 involves: (1) drawing k bootstrap samples

from the complete dataset; (2) fitting one tree for every one of them, with random features selection; (3) determining in which leaf each observation in $z_{j,mis}$ ends up according to the all of the k trees. Subsequently, for 2, the MI-RANF algorithm takes all donors from the k trees and randomly samples one imputation for each $z_{j,mis}$.

For greater details on the algorithms, the reader may consult algorithm A.1 in (Doove et al., 2014, p. 103, appendix B). The programming of the algorithm was heavily inspired by the *impute.mice.rf()* function in the R package *mice*.

2.1.7 MICE optimal model (MI-OP)

MI-OP is an ideal specification of the MICE algorithm, using as building block a univariate Bayesian imputation under the normal linear model, that includes as predictors in the imputation models the following types of variables:

1. all the variables in the complete-data analysis models;
2. all the variables that are related to the non-response;
3. all the variables are correlated with the target variables.

Following these criteria is one of the most commonly recommended strategies to deal with a large number of possible predictors for the imputation models (van Buuren, 2012, p. 168). In practice, researchers can never be sure requirement 2 is fulfilled, as there is no way to know exactly which variables are responsible for missingness. The MI-OP approach used here remains *ideal* in the sense that it is not entirely applicable in practice, but it offers an optimal benchmark point.

2.2 Single data strategies

missForest High dimensional imputation is often addressed with single imputation techniques. Most research on high-dimensional data imputation has focused on applications for DNA genetics data where the goal is to allow the use of large datasets for high-dimensional predictive algorithms, rather than inferential analysis. For this reason, a variety of single imputation machine learning algorithms have been proposed and compared (de Andrade Silva and Hruschka, 2009; Stekhoven and Bühlmann, 2011).

In this study, we consider the missForest imputation method proposed by Stekhoven and Bühlmann (2011), which is a popular non-parametric imputation approach (which does not suffer from the problem of unidentified imputation models) that can accommodate for mixed data type of the missing variables, and has been robustly implemented in a popular R-package (Stekhoven, 2013). The approach consists of an iterative imputation that first trains a RF on observed values, and then uses it to predict the missing values.

This is a single imputation method we do not expect it will perform well for inferential tasks, at least compared to the other high dimensional MI methods discussed here. Nevertheless, it is interesting to include this method as reference.

Complete Case Analysis Most data analysis software either ignore the presence of missing values or default to list wise deletion: only complete cases are used for the analysis (R Core Team, 2020; pandas development team, 2020). As a default behaviour of most analysis tools, Complete Cases Analysis remains one of the most popular missing data treatments in social sciences, despite its renown flaws ((Rubin, 1987, p. 8), (van Buuren, 2012, p. 9), Baraldi and Enders (2010)). Therefore, this method was included as a reference in this study.

Gold Standard Finally, the substantive models are also fitted to the underlying fully observed data. Results obtained in this fashion are referred to here and in the results tables as the Gold Standard method. They represent the counterfactual analysis that would have been performed if there had been no missing data.

3 Simulation Studies

The simulation study was broken up in two separate experiments: (1) the first was used to define a baseline comparison of the methods on multivariate normal data in both high and low dimensional conditions; (2) the second was used to assess the performance of the methods in the presence of a latent structure, in order to reflect the fundamental structure of social survey data.

3.1 Simulation Study Procedure

To assess the statistical validity of the different imputation methods we have repeated the following steps 1000 times for each experiment:

1. Data generation: A data matrix $\boldsymbol{Z}_{n\times p}$ was generated according to an experiment specific data generating model (e.g. multivariate normal model, confirmatory factor analysis), the characteristics of which depend on experimental conditions described below.

2. Missing data imposition: Missing values were imposed on a given number of target variables in $\boldsymbol{Z}_{n\times p}$, according to some response model.

3. Imputations: Each method described in section 2 to deal with missing values was used for imputation.

4. Analysis: Different analysis models were fitted to the differently treated data. Parameters estimates were pooled across the differently imputed datasets, for the MI methods, and stored along with the estimates obtained with single imputation methods and complete case analysis.

The code to Run the simulation was written in the R statistical programming language (version 4.0.3). All experiments were run using a 2.6 GHz Intel Xeon(R) Gold 6126 processor, 523780 MB of Memory. The operating system was Windows Server 2012 R2.

Computations were run in parallel across 30 cores. Parallel computing was implemented using the R package *parallel* and to ensure replicability of the findings seeds were set using the method by L'ecuyer et al. (2002) implemented in the R package *rlecuyer*. Code to run the studies can be found at `https://github.com/EdoardoCostantini/imputeHD-comp`.

In the following, each step of the simulation procedure is described in details for both experiments.

3.1.1 Step 1: Data generations

Experiment 1 The $\boldsymbol{Z}_{n\times p}$ data matrix in step 1 was generated by drawing from a multivariate normal distribution centered around 0 with a covariance matrix $\boldsymbol{\Sigma}_0$, with diagonal elements (variances) equal to 1. The off-diagonal elements of $\boldsymbol{\Sigma}_0$ were used to define three blocks of variables: the first five variables were highly correlated among themselves ($\rho = .6$); variables 6 to 10 were slightly correlated with variables in block 1 and among themselves ($\rho = .3$), and all the remaining $p - 10$ variables were uncorrelated. Items were rescaled to have mean of 5.

Experiment 2 The observed data $\boldsymbol{Z}_{n\times p}$ was created based on a Confirmatory Factor Analysis model. Each of l latent variables was assumed to be measured by 5 items, for a total of $p = 5 \times l$ columns in $\boldsymbol{Z}$. Values on the observed items for the i-th observation were obtained with the following measurement model:

$$\boldsymbol{z}_i = \boldsymbol{\Lambda}\boldsymbol{\xi}_{i.} + \boldsymbol{\delta}_{i.} \tag{9}$$

where $\boldsymbol{z}_i$ is a vector of $5 \times l$ observed items scores, for observations $i = 1, ..., n$; $\boldsymbol{\Lambda}$ is the matrix of factor loadings; $\boldsymbol{\xi}_{i.}$ is a vector of scores on the latent variables for observation i; and $\boldsymbol{\delta_{i.}}$ is a vector of uncorrelated multivariate normal measurement errors. For notation and model specification the interested reader may refer to **?**. All items are centered around a mean value of 5.

The latent scores in $\boldsymbol{\xi}_{i.}$ are sampled from a multivariate normal distribution centered around 0, and with a covariance matrix $\boldsymbol{\Psi}_0$, with diagonal elements equal to 1 and off-diagonal elements equal to correlation between latent factors. In particular, the first 4 latent variables are highly correlated ($\rho = .6$), the second block of 4 latent variables are somewhat correlated ($\rho = .3$), while the remaining $l - 8$ latent variables are uncorrelated.

The matrix $\boldsymbol{\Lambda}$ defines a simple latent structure where each item loads on only 1 factor (5 items for each latent variable). Both the item and latent factor variances are set to 1, $var(x_i) = 1$ and $\Psi_{ii} = 1$, so that the measurement error is defined as $var(\delta) = 1 - \lambda^2 1$. This specification allows factor loadings λ_{ij}, with $i = 1, ..., n$ and $j = 1, ..., l$, to be defined as standardized values that range between 0 and 1. If all values in $\boldsymbol{\Lambda}$ are 0s, there is no latent structure and items are simply drawn from multivariate distribution centered around the item means with covariance matrix $\boldsymbol{\Psi}_0$. If all values in $\boldsymbol{\Lambda}$ are 1s, there is a *prefect* latent structure, meaning that items exactly measure the latent constructs. The exact values for the latent factors are drawn for each repetition from a uniform distribution between a lower and upper bound, b_l and b_u, that are condition-specific (see below).

Conditions The data generating mechanisms described above were specified according to different conditions which are described below. Table 1 summarizes these conditions.

Experiment 1 Two experimental factors were considered: p, the number of columns in the dataset which are all fed to the imputation algorithms, taking value 50 or 500; and pm, the target proportion of *per* variable missing cases, taking value 0.1 or 0.3 The sample size n was set to 200 in all conditions.

Experiment 2 The dimensionality of the data was controlled based on the number of latent variables l. Two values were used for this factor: 10 and 100. In all conditions, 5 items were generated as measures for each latent variable, making conditions with $l = 10$ low dimensional conditions, with 50 total predictors and a constant sample size of 200 observations, and conditions with $l = 100$ high dimensional ones, with data matrices of dimensionality 200×500. The proportion of missing values was defined again as a fixed experimental factor with two levels: 0.1 or 0.3.

In experiment 2, we also defined the factor loadings λ_{ij} as a 2-level random experimental factor. The data generation step used factor loadings drawn from a uniform distribution defined between either 0.5 and 0.6, or or 0.9 and 0.97.

condition	n	p	l	pm	λ range
Experiment 1					
1	200	50	-	.1	-
2	200	500	-	.1	-
3	200	50	-	.3	-
4	200	500	-	.3	-
Experiment 2					
1	200	50	10	0.1	[.9, .97]
2	200	500	100	0.1	[.9, .97]
3	200	50	10	0.3	[.9, .97]
4	200	500	100	0.3	[.9, .97]
5	200	50	10	0.1	[.5, .6]
6	200	500	100	0.1	[.5, .6]
7	200	50	10	0.3	[.5, .6]
8	200	500	100	0.3	[.5, .6]

Table 1: Summary of conditions for experiment 1 and 2

3.1.2 Step 2: Missing data imposition

The non-response mechanism was modelled as a logistic regression:

$$logit[p(x_{i,t} = miss|X)] = \theta_0 + \tilde{X}_i\boldsymbol{\theta} \tag{10}$$

where $x_{i,t}$ is the response of variable t for the i-th subject, θ_0 is the intercept parameter, $\tilde{X}_i$ is a vector of responses for the i-th individual to the set of predictors involved in the missing data mechanism, and $\boldsymbol{\theta}$ is the vector of slope parameters. The probability of an individual non response was computed as the inverse of the logit function as

$$p_{miss} = p(x_{i,t} = miss|X) = \frac{exp(\theta_0 + \tilde{X}\boldsymbol{\theta})}{1 + exp(\theta_0 + \tilde{X}\boldsymbol{\theta})} \tag{11}$$

Finally, an n dimensional response vector was sampled from a binomial distribution $b(n, p_{miss})$. The value of θ_0 was chosen with an optimization algorithm that minimized the difference between a target proportion of missing values and its actual value.

Experiment 1 Six variables were chosen as target of missing data imposition: three variables in the block of highly correlated variables and three in the block of lowly correlated variables (x_t with $t = 1, 2, 3, 6, 7, 8$). Item non-response was imposed following equation (10) with 4 variables included in $\tilde{X}$: two fully observed variables from the highly correlated variables and two from the lowly correlated group of variables (x_r with $r = 4, 5, 9, 10$).

The choice of predictors in $\tilde{X}$ is important to allow imputations under MAR: the probability of observing a response for a target variable did not depend on the variable itself, to avoid imputation under Missing Not At Random; and, as all features in the data are included in the MI procedures, the predictors in $\tilde{X}$ are always allowed to be part of the imputation models.

Experiment 2 Item non-response was imposed on the 10 items measuring two highly correlated latent variables ($l = 1, 2$) using the other two highly correlated latent variables ($l = 3, 4$) as predictors in response model (10).

3.1.3 Step 3: Imputation

Missing values are dealt with according to all the methods described in section 2. Here, we describe some key details in the algorithms specifications.

Convergence For both experiments, convergence of the imputations was assessed in a preprocessing step. Before running the actual simulation studies 10 datasets were generated according to each experimental set up. Missing values in each dataset were imputed by running 5 parallel imputation chains for each Multiple Imputation method. Convergence was checked by plotting the mean of the imputed values for each variable in each stream, against the iteration number.

This procedure was performed only for the condition with larger proportion of missing values and larger number of predictors. Being this the most challenging condition for the imputation task, the decisions made for this condition were applied to the simpler ones.

In each parallel run, all the MI algorithms except run for 250 iterations. This number should be more then enough to check convergence of the MICE algorithms as in practice MICE requires a much lower number of iterations that more traditional MCMC algorithms (Van Buuren, 2018, p. 126). After approximately 40 iterations, the patterns shown in these trace-plots were free of trend, and the variance within chains was approximately the same as the variance between chains. Hence, to run the actual simulation experiments, the algorithms were considered to converge after 50 iterations, after which 10 imputed data sets were obtained and used for the subsequent standard complete-data analysis and pooling.

It was expected that blasso would require more iterations to converge than the other methods: the MCMC algorithm has to sample the penalty parameter τ and the sparsity parameter ρ on top of all the other parameters of the imputation models and the missing values replacements. Hence, to check convergence, each of the 5 parallel runs of the Blasso imputation algorithm was run with a total of 2000 iterations. For most variables and data repetitions, 500 iterations were more than enough to reach convergence of the imputations. However, some exceptions required more than 1250 iterations to be free of trends. Hence, in the single chain run used to obtain the simulation results, the 10 datasets used for complete-data analysis were kept after 1950 iterations of the blasso algorithm. Although this might seem a large number, blasso is a quite fast algorithm that allows this large number of iterations, at least in our set up.

Tuning penalty parameters The ridge penalty used in the bridge algorithm is fixed across iterations and its value needs to be decided beforehand by the imputer. After defining a grid of plausible values for the ridge penalty, the value used in the simulation was decided by means of cross-validation in a pre-processing phase. The grid was defined by eight values equally spaced in the range from 10^{-1} to 10^{-8}. After generating 100 datasets, bridge imputation was performed with each of the different penalty parameters and used to obtain 10 differently imputed datasets. For each data replication, the Fraction of Missing Information (fmi) (Savalei and Rhemtulla, 2012) associated with each parameter in the analysis models of interest (see next section for details) was computed and then averaged across repetitions. The mean of these average parameter fmis was used as a composite measure of fmi associated with each ridge penalty value. Finally, the penalty value with the smallest composite fmi was selected.

Both IURR and DURR can be specified with a variety of penalty parameters, from ridge to elastic net penalties. In this study we have specified the regularization as a lasso penalty, instead of a more elaborate elastic net, as the performance of the two were quite similar in (Zhao and Long, 2016; Deng et al., 2016) and the cross-validation is less computationally intensive for the former. A 10-fold cross-validation procedure is used at every iteration of DURR and IURR to choose the penalty parameter.

Blasso hyperparameters For blasso, in order to maintain consistency with previous research, the hyperparameters in 5, 6, and 7 were specified as in Zhao and Long (2016): $(a, b) = (0.1, 0.1)$, $(r, s) = (0.01, 0.01)$, and $(g, h) = (1, 1)$.

MI-PCA Number of components In the MI-PCA algorithm, enough components were extracted to explain 50% of the total variance in the data.

missForest iterations and number of trees To impute data with the single imputation random forest approach we provided the data with missing values to the function *missForest* in the homonymous R package. This function implements algorithm 1 in Stekhoven and Bühlmann (2011). The stopping criteria for the missForest algorithm is usually met under 10 iterations, but to make a conservative choice we fixed the maximum number of iterations to 20. Stekhoven and Bühlmann (2011) showed that increasing the number of trees grown in each forest has stagnating effects on the imputation error while increasing linearly the computation time. In their paper, the authors recommend growing 100 trees per forest, which offers a good compromise between imputation precision and computation time. Therefore, that is the value we used in this paper.

3.1.4 Step 4: Analysis

Experiment 1 The substantive model of interest in experiment 1 is a saturated model that estimates means, variances, and covariances of the six variables with missing values.

Experiment 2 The same saturated model is fitted to estimate the means, variances, and covariances of the *observed* items. Furthermore, the true Confirmatory Factor Analysis was also chosen to see how the factor loadings are recovered after imputation.

3.2 Comparison Criteria

After running the simulation procedure $R = 1000$ times, the R Gold Standard estimates, obtained by saving the model parameter estimates computed on the fully observed data, are averaged to define the true parameters values in each condition. The R estimates obtained by estimating the parameters of interest, after treating the missing values with all other methods, are used to compute the methods performance measures. In what follows, we describe the outcome measures that were considered.

Bias First, we used Percent Relative Bias (PRB) to quantify the bias introduced by the imputation procedures:

$$PRB = \frac{\bar{\hat{\theta}} - \theta}{\theta} \times 100 \tag{12}$$

where θ is the *true* value of the focal parameter (e.g. mean of item 1, variance of item 2) computed as $\sum_{r=1}^{R} \hat{\theta}_r^{GS}/R$, with $\hat{\theta}_r^{GS}$ being the Gold Standard parameter estimate for the r-th repetition. $\bar{\hat{\theta}}$ represents the focal parameter estimate under a given imputation method, averaged over the MCMC replications, computed as $\sum_{r=1}^{R} \hat{\theta}_r/R$, with $\hat{\theta}_r$ being the estimate obtained after using a given imputation approach in the r-th repetition. PRBs larger than 10% in absolute value are usually considered as extreme.

Confidence Intervals Coverage To assess the integrity of hypothesis testing, the Confidence Interval Coverage of the reference value was considered as

$$CIC = \frac{\sum_{r=1}^{R} I(\hat{\theta} \in \widehat{CI}_r)}{R} \tag{13}$$

where θ and $\widehat{CI}_r$ are, respectively, the parameter estimate and confidence interval of the focal parameter in a given repetition, and $I(.)$ is the indicator function that returns 1 if the argument is true and 0 otherwise.

CICs below .9 are considered problematic for 95% confidence intervals (Van Buuren, 2018, p. 52) as they imply inflated Type I error rates. A high coverage (e.g., .99) may indicate confidence intervals that are too wide, implying that the imputation method leads to more conservative inferential conclusions, and in this sense it is less worrisome than lower than nominal coverage. Therefore, Confidence Intervals were considered to show severe under-coverage (over-coverage) if they are below 0.9 (above .99).

In the present work, we followed Burton et al. (2006) and considered as problematic CI coverage rates outside of two Standard Errors (SE) of the nominal coverage probability (p). The standard error of nominal coverage is defined as $SE(p) = \sqrt{p(1-p)/R}$, with $p = .95$. For $R = 1000$, we considered CI coverages outside the (0.94, .96) were considered showing mild signs of deviation from nominal coverage.

3.3 Results

3.3.1 Experiment 1

Figure 1 and 2 report the Percentage Relative Bias and Confidence Interval Coverage, respectively, for each parameter estimate in the saturated model described above. For each method, single horizontal lines, representing the PRB (or CIC) of a single parameter estimate, combine and form larger horizontal bars giving an aggregate account of how the method performs across many multiple variables with missing values.

Means Focusing first on the item means (top rows), all methods achieved a bias that is smaller than the 10% threshold for all item means, in all conditions. Looking at relative performances, in all conditions, MI-OP, which is not applicable in applied research, provided negligible bias, and IURR and MI-PCA resulted in the smallest estimation bias among the other methods.

In the conditions with high proportion of missing values (condition 3 and 4), all methods showed some signs of under-coverage of the 95% CI, with CIC outside of the interval (.94, .96). The only exceptions to the trend were MI-OP, MI-PCA, and Bridge which showed little deviation from nominal coverage.

Variances Moving to the item variances (central rows), IURR, Blasso, and the tree based methods gave the lowest biases across all conditions, even in the most challenging one (condition 4). These low biases were mostly paired with low deviations from nominal coverage, except for condition 4 where IURR and the tree based methods showed clear signs of under-coverage of the true item variances ($CIC << .9$). Apart from MI-OP, Blasso was the method with best coverage in this final condition.

Directly using regularized regression within the imputation models (DURR) showed poor performances with regards to variances: in all conditions but the first, it led to large (negative) bias, with a PRB as large as 10% in the last condition, accompanied by clear signs of CI under-coverage. Bridge was the only MI method showing larger item variance bias than DURR, in all the high-dimensional conditions (2 and 4): imputing the data with Bridge led to all item variance estimates showing a bias larger than 20% the size of the true value.

MI-PCA also showed poor performances with a noticeable positive bias in all conditions that became extreme in condition 4, where the PRBs exceeded 20%. This poor performance was reflected in extreme confidence interval under-coverage of the true item variances in the final experimental condition.

Single data imputation method missForest and complete case analysis led to substantial negative bias and CI under-coverage for all item variances, even in condition 1.

Covariances Finally, the third row in figure 1 shows the estimation bias for the 15 covariances between the 6 items with missing values. As covariances depend on two variables, recovering the correct estimates after imputation is inherently more difficult than with means and variances. This explains the generally worse performances reported in the figure.

Indirect Use of Regularized Regression (IURR) performed noticeably better than most other methods, with negligible negative bias and acceptable coverage for all covariances in conditions 1, 2 and 3, but it struggled with a large negative bias and extreme under-coverage for the majority of the 15 covariances in condition 4.

MI-PCA showed negligible negative biases of the covariance estimates in all conditions, performing as well as MI-OP in all but the last condition, where it still maintained acceptably low values of PRB ($PRB < 10\%$). Furthermore, MI-PCA showed virtually no deviation from nominal coverage, with a CIC pattern similar to that of MI-OP, in all but the last condition, where it manifested only mild *over*-coverage of the items covariances.

All other methods, including DURR, showed large biases (larger than 10% threshold in absolute value) in all but the first condition, with mild to extreme signs of under-coverage. Single data approaches, like missForest and CC, showed extreme bias and under-coverage of covariances between items with missing values, even in condition 1. Bridge was again showing acceptably low biases and coverage in the low dimensional conditions, and extremely large biases and low CI coverage in all the high dimensional conditions.

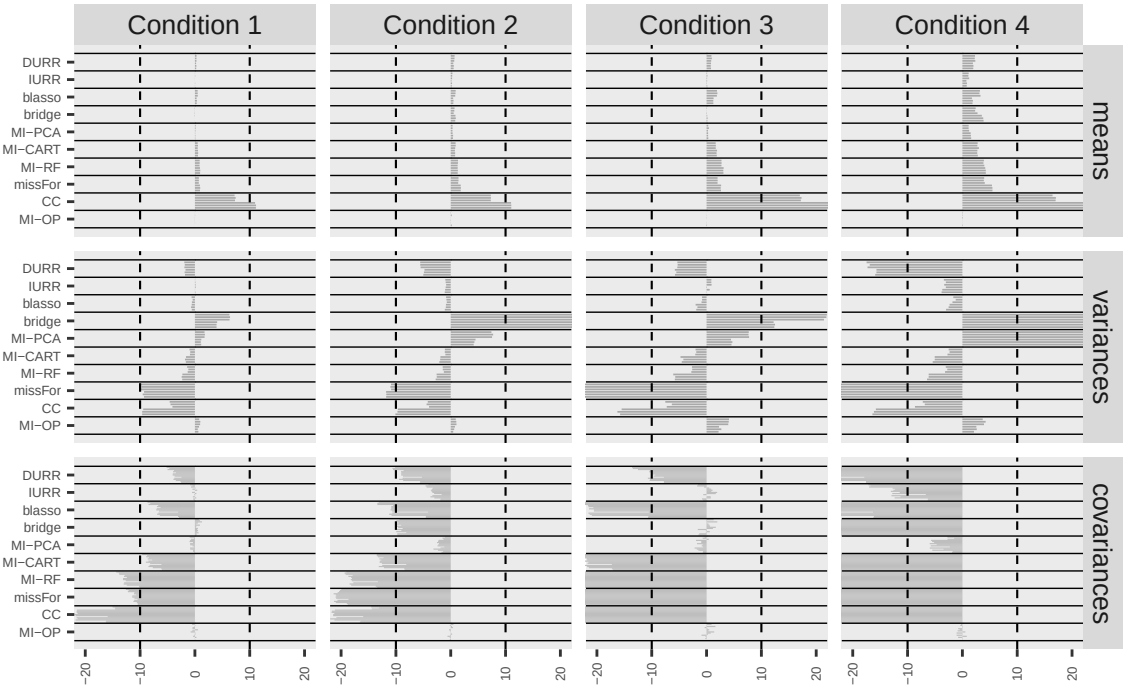

Figure 1: Percent Relative Bias (PRB) for item means, variances, and covariances. Within each panel, for every method, single horizontal lines report the PRB, for the same parameter, computed for each item (and pair of items) with missing values.

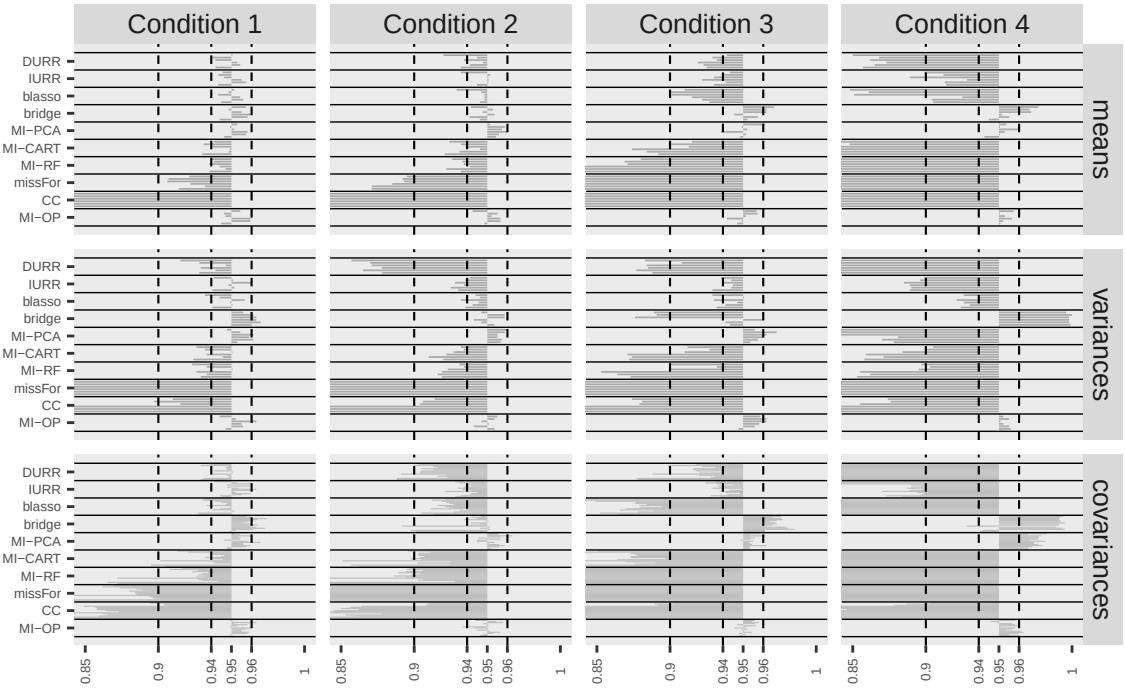

Figure 2: Confidence Interval Coverage (CIC) for item means, variances, and covariances. Within each panel, for every method, single horizontal lines report the CIC, for the same parameter, computed for each item (and pair of items) with missing values. The vertical lines reported are thresholds signposting deviations from nominal coverage that are worrisome. CIC below .9 represent extreme under-coverage, while the Burton et al. (2006) threshold (.93 and .97) represent milder deviations from nominal coverage.

3.3.2 Experiment 2

Figure 3 and 4 report the Percentage Relative Bias and Confidence Interval Coverage, respectively, of the estimated means, variances, and covariances of the 10 observed items with missing values in the first four conditions of experiment 2, the ones with high factor loadings (strong latent structure). For each method, single horizontal lines, representing the PRB (or CIC) of a single parameter estimate, combine and form larger horizontal bars giving an aggregate account of how the method performs across many parameters. Figure 10 and 11 in appendix reports the same values for the low factor loading conditions.

Means All methods provided unbiased estimates of the item means with PRBs that were almost 0 for all items. As the proportion of missing cases increased, in conditions 3 and 4, there was a slight increase in PRB values for all methods except IURR, bridge, and MI-PCA. However, only Complete Case analysis led to unacceptable bias in these scenarios. DURR, IURR and MI-PCA also showed little to no deviations from nominal coverage in all conditions, while Blasso, MI-CART, MI-RF, and Bridge showed important sings of under-coverage when the proportion of missing cases was high (conditions 3 and 4). missForest also led to extreme under coverage of the true values.

Variances All MI methods, except Bridge, showed acceptable bias levels for item variances estimates in all conditions, but the least biased estimates were obtained by MI-OP, IURR and MI-PCA. Confidence Intervals Coverage decreased as the proportion of missing cases increased (from condition 1 and 2 to conditions 3 and 4, respectively). Only IURR and MI-PCA maintained CICs mostly within the range .94 and .96, with the former showing slight signs of under-coverage and the latter tending toward over-coverage, while blasso and the MI tree-based methods showed sings of mild and extreme under-coverage (CIC $< 90\%$), respectively.

The large positive bias (and low CIC) for the item variances that afflicted MI-PCA in the multivariate-normal set up (figures 1 and 2) is not present in figures 3 and 4. However, that pattern reappeared when the latent structure was weak, as can be seen in figure 10 and 11 in the appendix (conditions 5 to 8, factor loadings between .5 and .8).

Single data approaches, missForest and CC, showed again extreme (negative) bias and CI under-coverage in almost all conditions.

Covariances IURR and DURR showed acceptable biases (PRBs below 10% in absolute value) in conditions 1 to 4, but the large negative covariance bias and extreme low coverage shown in the first experiment (see figures 1 and 2) reappeared when the latent structure was weak, as can be seen in figure 10 and 11 in the appendix (conditions 5 to 8, factor loadings between .5 and .6).

The other methods performed exactly as in experiment 1: the MI-PCA approach resulted in the lowest bias and deviation from nominal coverage for the covariances of the observed items; all other methods led to large negative biases and mild-to-extreme under-coverage for all the covariances, in all conditions.

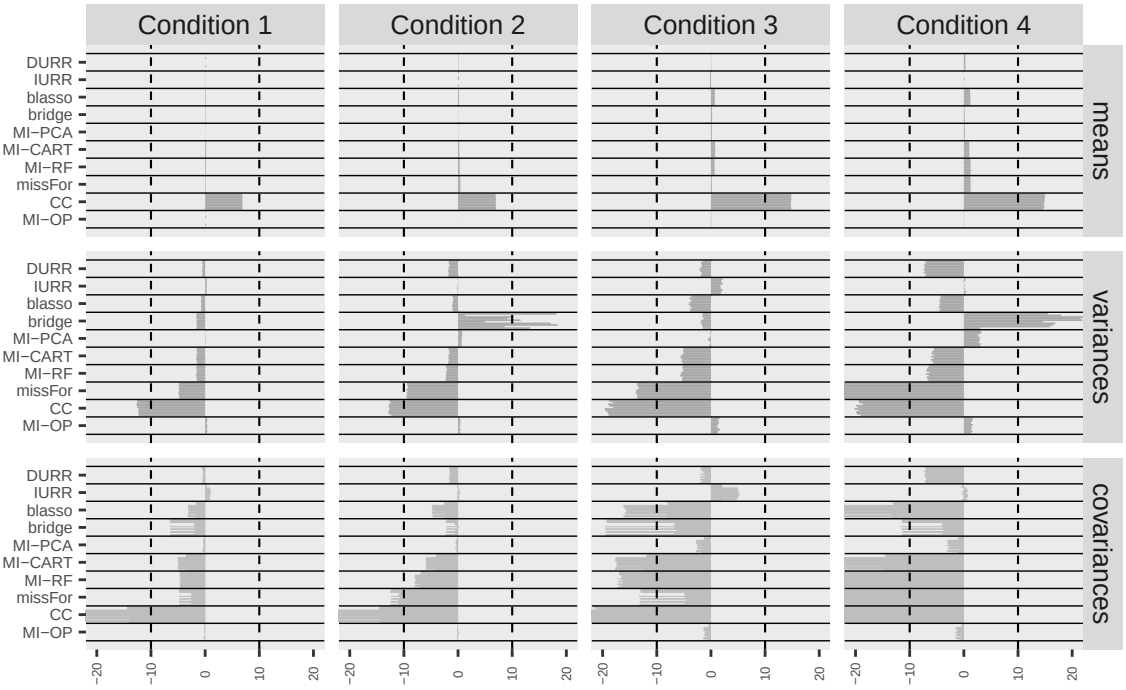

Figure 3: PRBs for the means, variances and covariances (PRB) for condition 1 to 4. Within each panel, for every method, single horizontal lines report the PRB, for the same parameter, computed for each item (and pair of items) with missing values.

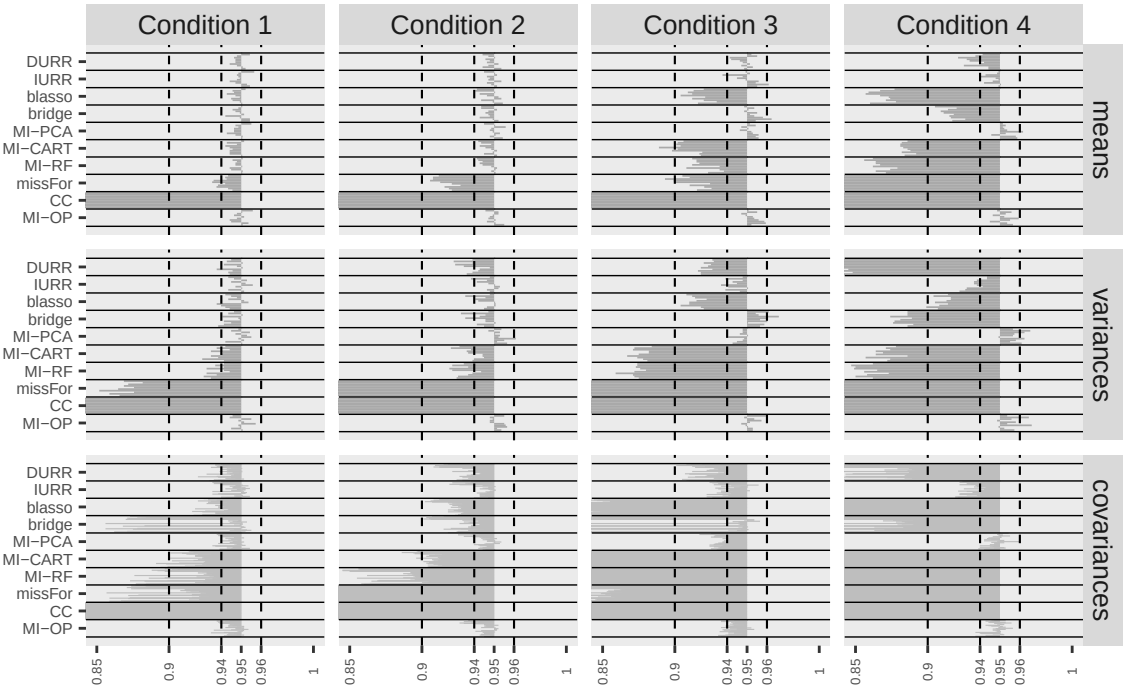

Figure 4: CIC for the means, variances, and covariances for condition 1 to 4. Within each panel, for every method, single horizontal lines report the CIC, for the same parameter, computed for each item (and pair of items) with missing values.

Factor Loadings Figures 5 shows the PRB values for all the factor loadings estimated by the Confirmatory Factor Analysis described above. Most MI-Methods provided acceptably low bias for these estimates in all conditions except the one with both large proportion of missing values and high dimensional input data matrix (condition 4 and 8).

MI-OP, IURR, and MI-PCA outperformed all other methods giving virtually unbiased estimates of the factor loadings in all conditions. In particular, MI-PCA outperformed IURR when factor loadings were low (panel b, conditions 5 to 8), maintaining inconsequential biases even when data is high-dimensional and the proportion of missing values was high.

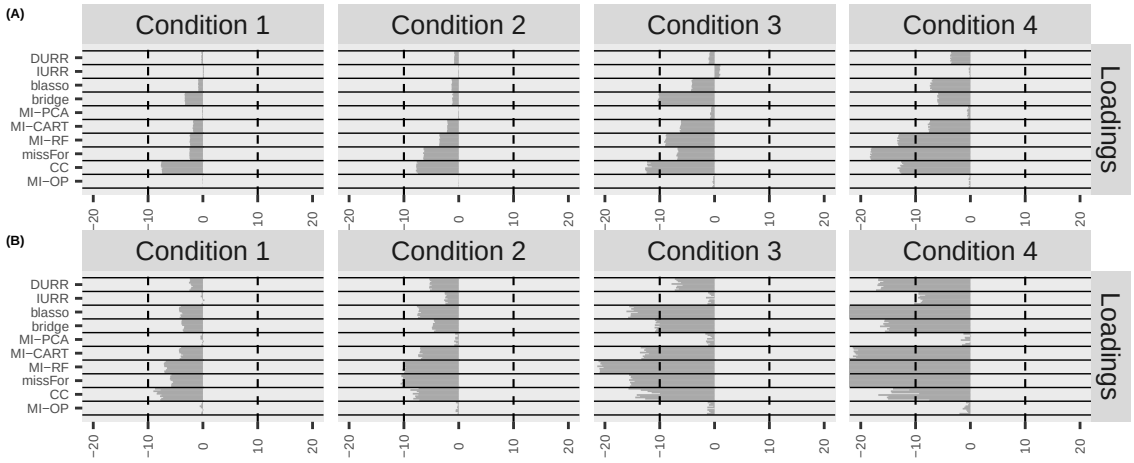

Figure 5: Percent Relative Bias (PRB) for the factor loadings in conditions 1 to 4 (panel A) and conditions 5 to 8 (panel B). Within each panel, for every method, single horizontal lines report the PRB of the factor loading estimation for each item with missing values.

4 Resampling Study

To test the ecological validity of findings in experiment 1 and 2 we designed a resampling study based on European Values Survey (EVS) data.

EVS is a large scale, cross-national survey on human values administered in almost 50 countries across Europe. It covers a wide range of human values and topics such as family, work, environment, perceptions of life, politics and society, religion and morality, national identity. It is a high quality survey widely used for comparative studies between European countries. Furthermore, it is accessible free of charge and it represents the type of data social scientist regularly work with.

Variables in the EVS data are not generated artificially from continuous normal distributions, or predefined Factor Analysis models, but are discrete numerical and categorical items following a variety of distributions. By using data gathered for an actual survey, we could study whether the relative performances of the imputation methods, displayed in the simulation studies, changed when deployed for real data research.

It is useful to think of a research scenario to go along the resampling study. Imagine a researcher that wants to test some hypothesis derived from sociological theory and decides to analyse a large comparative social survey to address their research question. The data has missing values and decisions must be made on how to treat them. The researcher wants to use an imputation procedure that discards as little information as possible while making as little arbitrary decisions as possible as to which predictors to include in the imputation models. This researcher will inevitably have to decide which imputation method to use and how to specify it.

4.1 Resampling Study Procedure

The resampling study followed a similar strategy to that used in for the simulations. To assess the statistical validity of the different imputation methods we repeated the following steps 1000 times ($R = 1000$):

1. Data generation: A bootstrap sample $\boldsymbol{Z}^*$ is generated by sampling with replacement n observations from a pre-processed EVS data-matrix. Part of the pre-processing step is the imputation of the extant missing data to obtain a pseudo-fully observed input data matrix so that $\boldsymbol{Z}^*$ has no missing values;

2. Missing data imposition: Missing values are imposed on a given number of target variables in $\boldsymbol{Z}^*$, according to some response model (see below), and $\boldsymbol{Z}^*_{miss}$ is obtained;

3. Imputations: Each method described in section 2 is used to deal with missing values in $\boldsymbol{Z}^*_{miss}$.

4. Analysis: Two analysis models are fitted to the differently imputed data. Their parameter estimates are pooled across the differently imputed datasets, for the MI methods, and stored along with the estimates obtained after using single imputation methods, complete case analysis, and the Gold Standard method.

The average estimate, over the R repetitions, obtained with the Gold Standard approach were considered as “true" reference values of the parameters in the analysis models. The R estimates obtained with all other methods were used to obtain performance measures for each imputation method using the same criteria described for study 1 and 2 (see 3.2).

4.1.1 Data preparation

For this study we used the third pre-release of the 2017 wave of EVS data (EVS, 2020). The original dataset contained 55,000 observations in 34 countries. We selected only the four founding countries of the European Union included in the dataset (France, Germany, Italy, and the Netherlands) and excluded all columns of the data that were either duplicated information (recoded versions of other variables), or meta data (e.g. time of interview, mode of data collection).

All originally missing values were filled in with a run of a single imputation predictive mean matching (PMM) algorithm to obtain a pseudo fully-observed dataset. PMM was chosen for the task as it is a flexible imputation method that maintains the distributional characteristics of the original data. Bias and uncertainty introduced by this procedure is not relevant for the present study as the data matrix obtained after the single PMM run is treated as the population data.

At the end of this data cleaning process, we obtained a fully-observed dataset of 8045 observations (n), across 4 countries, and 243 variables (p).

Conditions There were only two conditions for the resampling study: low and high dimensional data imputation. As the number of predictors in the data is fixed ($p = 243$), the dimensionality of the data is changed by defining different sizes for the sample taken from the pseudo-fully observed data in step 1. We chose two values for n, namely 1000 and 300, corresponding to the low and high dimensional condition.

4.1.2 Analysis models

To define plausible analysis models we searched for analysis models that have been used in published articles testing sociological theories on the EVS data. The search was performed by screening the repository of publications using EVS data available on the EVS website.

As a results, we defined two linear regression models, model 1 and 2, of the same form:

$$y = \beta_0 + \beta_1 x_1 + \boldsymbol{\beta}_{-1}\boldsymbol{X}_{-1} \tag{14}$$

where a dependent variable y is regressed on a variable of interest x_1 and a set of control variables $\boldsymbol{X}_{-1}$. In this scenario, β_1 is a focal parameter that a researcher wants to use to test some hypothesis.

The first version of linear model 14, model 1, was inspired by Köneke (2014): $y^{(1)}$, its dependent variable, was a 10-point EVS item measuring euthanasia acceptance ('Can this always be justified, never be justified, or something in between?'); the predictor of interest $x_1^{(1)}$ was a 4-point item measuring the self-reported importance of religion in one's life; the matrix of covariates $\boldsymbol{X}_{-1}^{(1)}$ contains a selection of control variables (trust in the health care system, trust in the state, trust in the press, country, sex, age, education, and religious denomination).

This model represents a plausible analysis a researcher would perform to test an hypothesis regarding the effect of religiosity on the acceptance of end-of-life treatments.

Model 2, the second version of the linear model in equation 14, was inspired by Immerzeel et al. (2015). The dependent variable $y^{(2)}$ was an harmonized variable constructed by EVS to describe the respondents' tendency to vote left or right wing parties, expressed on a 10-point left-to-right continuum. The predictor of interest $x_1^{(2)}$ was a composite mean scale measuring respondents attitudes toward immigrants and immigration ('nativist attitudes scale'). The scale was obtained by taking the average of respondents expressed agreement, on a scale from 1 to 10, to three items: 'immigrants take jobs away from natives', 'immigrants increase crime problems', and 'immigrants are a strain on welfare system'. The control variables used were: attitudes toward low and order, attitudes toward authoritarianism, interest in politics, level of political activity, country, sex, age, education, employment status, socio-economic status, importance of religion in life, religious denomination, and the size of town where interview was conducted.

A researcher might fit this model and look at $\beta_1^{(2)}$, the 'nativist attitude' regression coefficient, value and standard error to test an hypothesis regarding the effect of xenophobia on voting tendencies.

4.1.3 Missing data imposition

Missing data were imposed on 6 variables according to the same strategy described in 3.1.2. The variables target of missing value imposition were $y^{(1)}$ and $y^{(2)}$, the two dependent variables in model 1 and 2; religiosity ($x_1^{(1)}$, focal and control variable in model 1 and 2 respectively), and the three items making up the "nativist attitudes" scale (focal predictor $x_1^{(2)}$ in the second model).

The response model form is the same as in equation 10 and three variables were included in $\tilde{X}$: age, education, and an item measuring trust in new people. These aspects may plausibly influence response tendencies in participants: older people usually have higher item non-response rates than younger people; lower educated people tend to have higher item non-response rates than higher educated people; people with less trust in strangers are assumed to have higher item non-response tendency as they are likely to withhold more information from the interviewer (a stranger).

4.1.4 Imputation

Missing values were dealt with according to all the methods described in section 2. Here, we describe some key details in the algorithms specifications.

Convergence As for the simulation studies, convergence of the imputations was assessed in a pre-processing step. Before running the actual simulation studies 10 datasets were generated according to each experimental set up. Missing values in each dataset were imputed by running 5 parallel imputation chains for each Multiple Imputation method. Convergence was checked by plotting the mean of the imputed values for each variable in each stream, against the iteration number. This procedure was performed only for condition 2, the high-dimensional one, under the assumption that convergence would be faster in the low-dimensional set up.

After the convergence check, we decided to run the algorithms for the regular experiment run for 60 iterations before saving the multiply imputed datasets, although most methods achieved convergence well before that number of iterations.

Tuning penalty parameters The ridge penalties used in the bridge algorithm were chosen with the same cross-validation procedure described for experiment 1 and 2.

IURR and DURR used a lasso formulation of the frequentist regularized regression and a 10-fold cross-validation procedure was performed at every iteration to choose the penalty parameter.

Blasso hyper-parameters In order to maintain consistency with previous research, the hyper-parameters in 5, 6, and 7 were specified as in Zhao and Long (2016): $(a, b) = (0.1, 0.1)$, $(r, s) = (0.01, 0.01)$, and $(g, h) = (1, 1)$.

MI-PCA Number of components In the MI-PCA algorithm, enough components were extracted to explain 50% of the total variance in the data. This choice is somewhat arbitrary and it is the product of the current lack research on what are the best solution for the imputation context.

missForest iterations and number of trees As in the simulation studies, the maximum number of iterations was set to 20 and the number of trees for the random forest was set to 100.

4.2 Results

4.2.1 Single Parameter of interest

Figures 6 and 7 report the PRB and the CIC for parameters $\beta_1^{(1)}$ and $\beta_1^{(2)}$, the focal regression coefficients in the two models.

Most of the MI methods resulted in negligible biases ($|PRB| < 10\%$) for both parameters in all conditions. The only two exceptions were bridge and MI-RF: the former was very competitive in condition 1, the low dimensional one, but led to extreme bias and over-coverage in the high dimensional condition for $\beta_1^{(2)}$; the latter provided the largest PRB and worst CI (under) coverage for the focal regression coefficients among the other MI methods, and it was consistently outperformed even by Complete Case analysis. missForest also displayed contained focal parameter biases.

DURR and IURR gave inconsequential biases for both parameters in all conditions, with PRBs that were often at least half in size as the ones obtained with the other methods, outperforming even MI-OP in the high dimensional condition for $\beta_1^{(2)}$.

For $\beta_1^{(2)}$ in model 2, both IURR and DURR remained fairly competitive in terms of coverage with CICs close to nominal levels, but the advantage they showed in terms of bias was not carried over to this criterion. Both MI-PCA and Blasso provided CICs either equal or closer to nominal than the ones obtained with IURR and DURR in almost all conditions.

As for $\beta_1^{(1)}$, all MI methods showed signs of under-coverage with CIC smaller than the threshold value .93. Although coverage of the true values was not great for any of the methods selected, the Gold Standard confidence intervals were also under-covering the true value of $\beta_1^{(1)}$. This was likely due to the right-skewed nature of the distribution of the dependent variable (euthanasia acceptance). Most MI imputation methods achieved coverages similar to that of the Gold Standard method, and more importantly their relative difference, compared to the GS coverages, was in line with what seen for $\beta_1^{(2)}$.

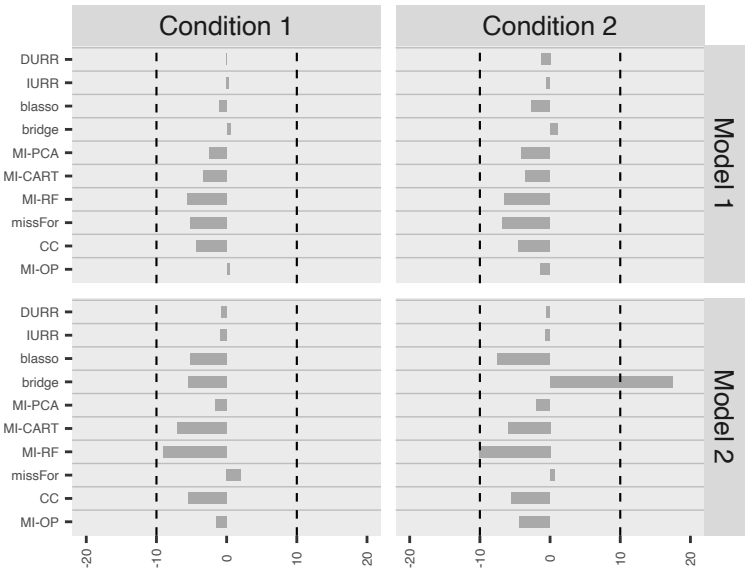

Figure 6: Bias for single parameter of interest in the two different models

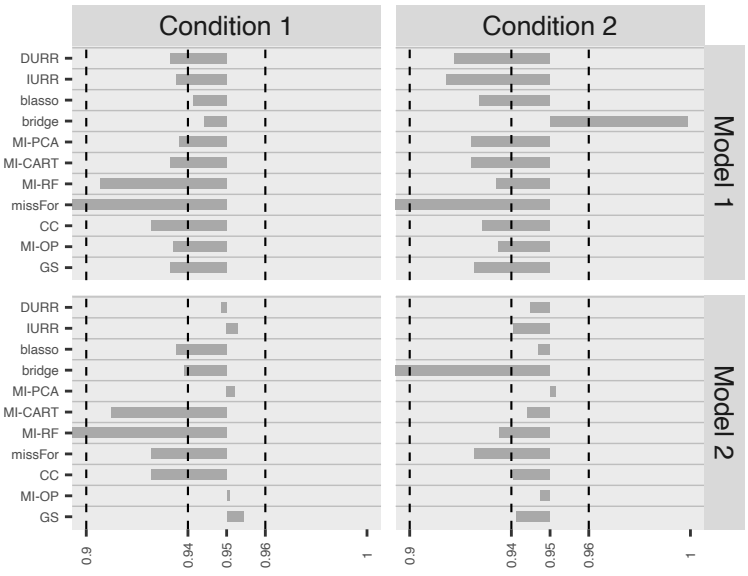

Figure 7: Confidence Interval Coverage for single parameter of interest in the two different models

4.2.2 Overall Model Parameter Assessment

Looking at bias and confidence interval coverage for a single parameter of interest can help assess the degree to which inferential conclusions, based on it, might change in applied research depending on the imputation method used. However, the fitted models were multiple regressions, and the estimate of all model parameters were influenced by the imputations, not only the regression coefficients of some variable of interest.

Therefore, it is important to observe the overall effects of the different methods on model estimation. Figure 8 reports the absolute values of the PRBs for each of model 2 parameter estimates, ordered by size, under each of the different missing data treatment considered. The figure shows, for each method, how many parameters exceed the 10% threshold, what is the largest and what is the smallest PRB achieved.

The regression coefficient for the country dummy code identifying the Netherlands has the largest bias across almost all methods and conditions. For this reason it is highlighted in the picture, along with the intercept and the focal regression coefficient.

MI-OP showed that even having perfect information regarding the missing data mechanism and data structure, results in some bias for certain estimates. Although the bias for the intercept and the focal parameter were negligible, around half of the estimates obtained after using this imputation method showed large biases ($|PRB| > 10\%$), and the largest bias was considerable (around 40%, in the low dimensional condition, and 20%, in the high-dimensional one).

In both the high- and low-dimensional condition, Multiple Imputation done with DURR, IURR, Blasso, and MI-CART, and single imputation done with missForest showed fairly similar overall patterns to MI-OP, with only slightly larger PRBs. MI-PCA and MI-RF also showed similar trends but they presented overall larger PRBs for those estimates exceeding the 10% threshold. However, none of these methods seemed to suffer from the increase in dimensionality.

Bridge demonstrated the same behaviour described in the simulation studies: it was a competitive method in low dimensional scenarios, but it was inadequate to deal with high-dimensional data imputation (all but one PRBs are larger than 100%).

Figure 9 reports the confidence interval coverages for each parameter estimate in model 2. When using MI-OP, for only two parameters CIC was showing a deviation from nominal coverage, with a slight tendency toward over-coverage. While DURR, IURR, MI-CART and MI-PCA maintained a similar coverage pattern to the oracle MI-OP approach, blasso, MI-RANF, and missforest were either over- or under-covering many more parameters. The behaviour showed by missForest was somewhat to be expected: as a single imputation approach, it underestimates uncertainty regarding values of the empty data cells, and tends to produce narrower confidence intervals.

Despite showing poor performances in terms of bias, Complete Case analysis manifested good coverage. However, this was a result of the smaller sample size used for fitting the analysis model, rather than a positive feature of the method: good CI coverage is desirable only as long as bias is contained.

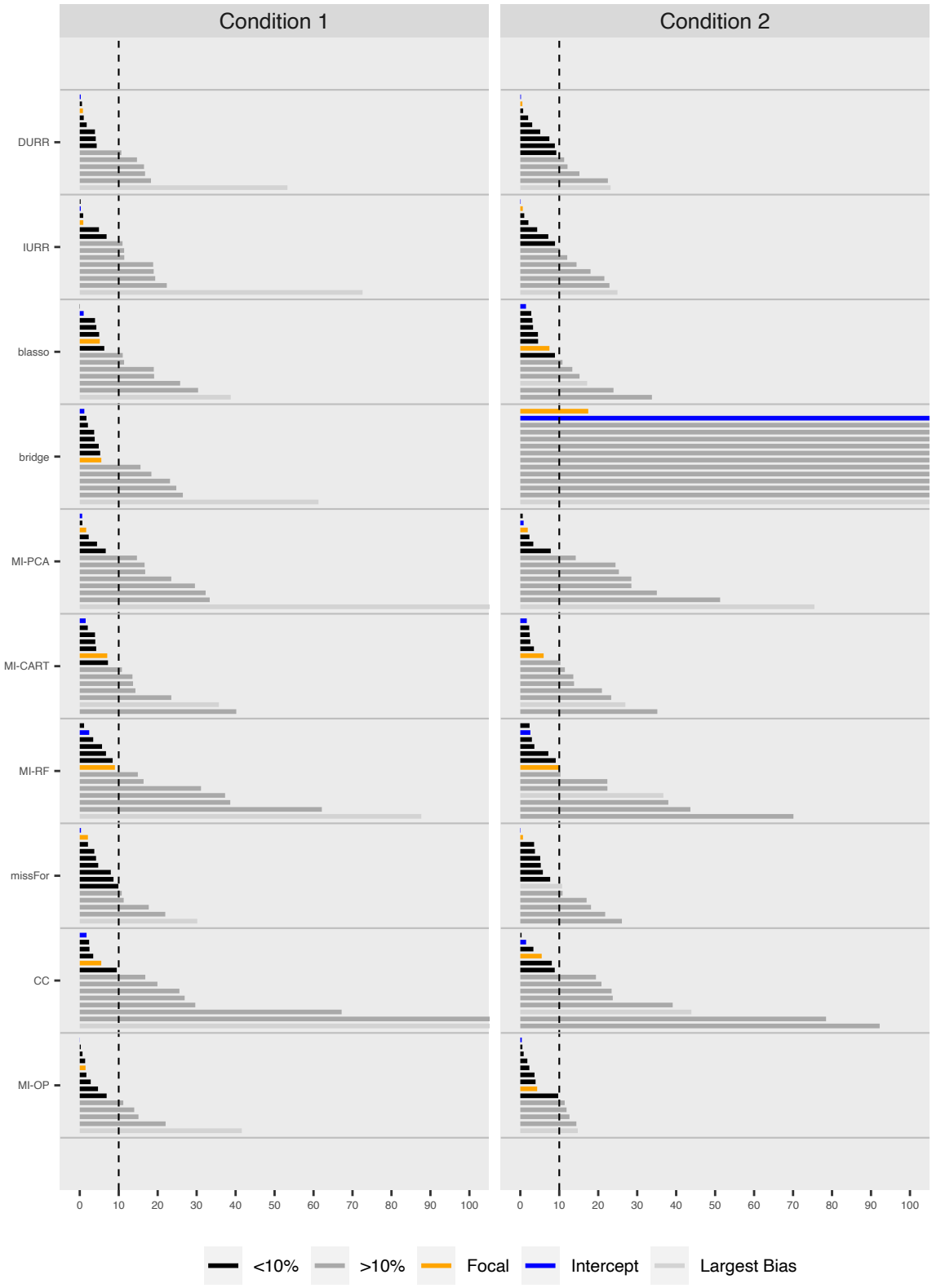

Figure 8: PRBs for all the model parameters in model 2. The order of the bars is based on the absolute value of the PRBs. The values for the intercept, the focal regression coefficient, and the regression coefficient with which most methods struggle (Largest Bias) are highlighted

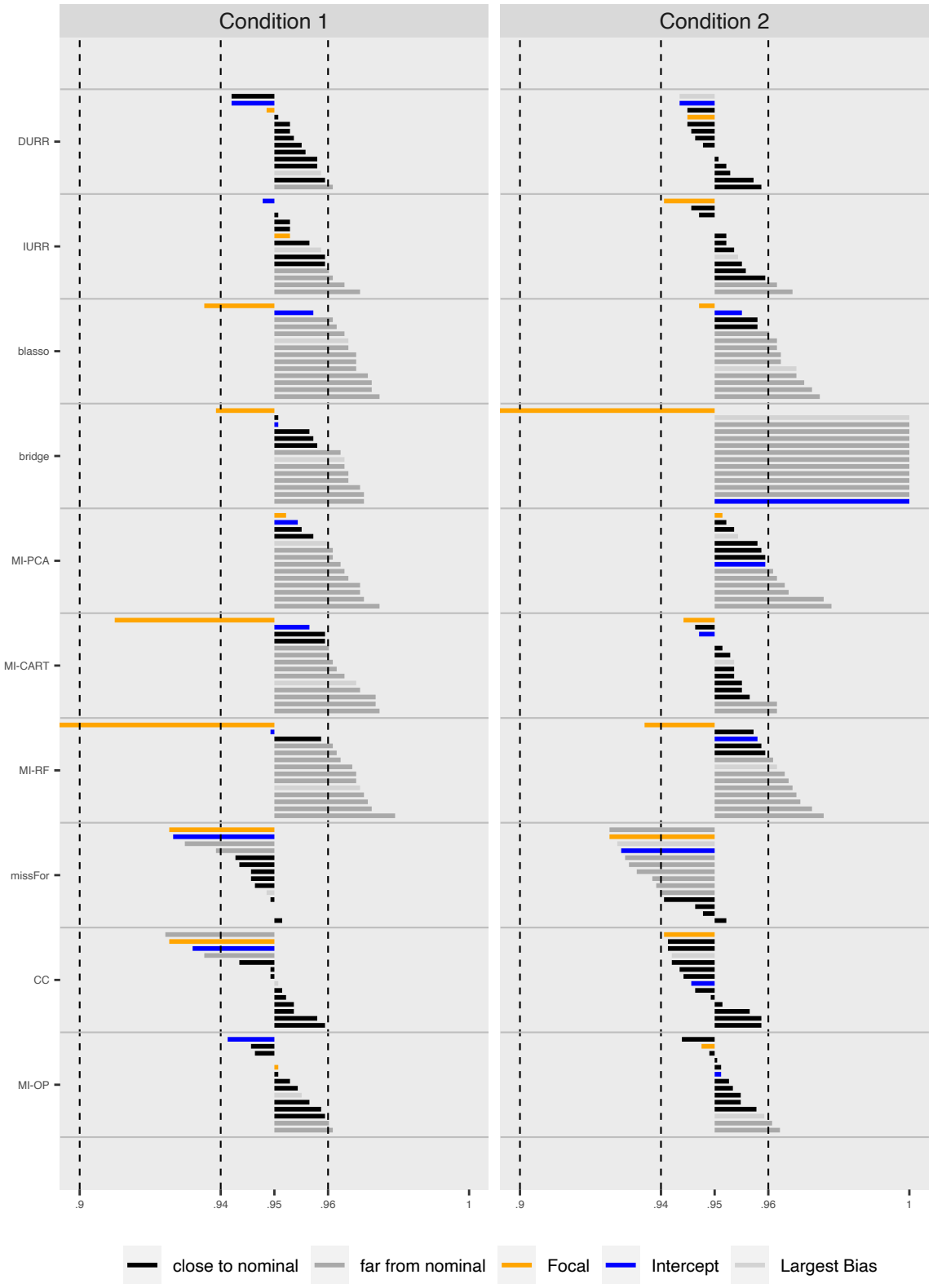

Figure 9: CIC for all model parameter in model 2. The values for the intercept, the focal regression coefficient, and the regression coefficient with which most methods struggle (Largest Bias) are highlighted

4.2.3 Imputation Time

Table 2 reports the average imputation time across the different methods. IURR and DURR were the most time consuming methods with imputation times above the hour, in our low dimensional conditions, versus imputation times of a minute or less for MI-PCA and Blasso imputation. In the high dimensional condition, IURR and DURR are not as time-intensive, but still require more then ten times the time of MI-PCA and blasso imputation.

	DURR	**IURR**	**blasso**	**bridge**	**MI-PCA**	**MI-CART**	**MI-RF**	**MI-OP**
Condition 1	73.20	75.90	1.40	8.10	0.60	4.00	11.30	2.20
Condition 2	6.10	9.70	0.50	3.20	0.40	1.40	4.70	1.90

Table 2: Average imputation time in minutes

5 Discussion

We investigated the relative performances of seven approaches to Multiple Imputation for general missing data patterns that do not require researchers making decisions on which variables to include in the imputation models. In this section, we summarize how the methods performed in our numerical set ups and comment on their strengths and weaknesses.

IURR and DURR Overall, both Direct and Indirect use of regularized regression within MICE (DURR and IURR) returned low bias and good CI coverage for item means and variances in the simulation studies, and for the regression coefficients in the resampling study. IURR in particular excelled with some of the smallest estimation biases for item means, variances and regression coefficients, while DURR struggled with large biases for item variances in the high-*pm*-high-dimensionality condition in experiment 1.

For item covariances, IURR delivered noticeably better performances than all other methods, except MI-PCA. In the high-*pm*-high-dimensionality conditions, most MI methods resulted in PRBs larger than 20% in size, and CICs well below 0.9, while the negative covariance estimation bias introduced by IURR in the simulation studies was just slightly larger than the 10% threshold and the CI coverage was just around 0.9.

The performances showed by DURR and IURR come at a large computational cost: in our resampling study set up, they took significantly more time to perform on average than all other methods.

MI-PCA Overall, MI-PCA showed low biases for item means and covariances in both simulation 1 and 2. It was the only method showing acceptably low bias and close-to-nominal CI coverage of the true covariance values in the most challenging conditions of experiment 1 and 2.

MI-PCA showed poor performances in terms of estimation bias of the item variances. In the high-*pm*-high-dimensionality condition of experiment 1, MI-PCA led to item variance PRBs larger than 20%. The bias for the item variances that afflicted MI-PCA in the multivariate-normal set up appeared to be related to the strength of the latent structure: when the latent structure was absent (experiment 1) or weak (experiment 2, conditions 5 to 8, factor loadings between .5 and .6) item variances were biased, especially in the high-dimensional conditions; when the latent structure was prominent (experiment 2, conditions 1 to 4), the variances were estimated with negligible bias, even in the high-dimensional conditions. One possible explanation for this is that when data comes from a CFA model with factor loadings close to 1 (experiment 2, conditions 1 to 4), variables measuring the

same latent construct become increasingly correlated, and as a result imputation becomes more accurate and the item variance estimates less biased.

MI-PCA performed acceptably in the resampling study, with sufficiently low biases and close to nominal CI coverage for the focal parameters and the overall model parameter assessment. However, the great recovery of bivariate relationships manifested in experiment 1 and 2 (low covariance bias) did not directly translate in particularly low biases for regression coefficients. PCA is a tool to find a low-dimensional representation of a data set summarizing in a few components as much unique variation on each dimension of the data as possible. As such, PC extraction is likely negatively affected by the nature of survey data. Items in surveys are usually discrete, and their possible values are only a few integer values in relatively small ranges. This key aspect might be the root of the different performances obtained by MI-PCA in the simulation and the resampling study.

Bridge In both the simulation and resampling study the use of a fixed ridge penalty within the imputation algorithm to facilitate the inversion of the observed data matrix manifested the same behaviour: the method was competitive when many predictors were included in the imputation model but the problem remained low dimensional, while it led to extreme bias, and consequently unacceptable confidence interval coverage, in all the high dimensional conditions.

Blasso Overall, Blasso showed good performances in terms of bias, keeping the absolute PRBs for item means and variances below 10% in the high-dimensional conditions of experiment 1 and 2. While PRBs were high for covariances in these experiments, blasso remained one of the top performer in the resampling study, where the overall pattern of regression coefficients PRBs was quite similar to that of MI-OP.

However, in terms of confidence interval coverage, blasso showed poor performances resulting in either CI under-coverage or CI over-coverage of true parameter values in almost all high-dimensional conditions, across the three different experimental set ups. Furthermore, blasso did not fair particularly well in allowing an unbiased recovery of the latent structure in our second simulation study, as the PRBs for factor loadings were the highest among the MI methods.

Using Hans (2010)'s Bayesian Lasso requires the specification of 6 hyper-parameters, which introduces more researcher degrees of freedom and demands more familiarity with Bayesian analysis. Although there are recommendations in published work on what values to use for these hyper-parameters, we have not investigated the sensibility of results to different values.

Alternative implementations of Bayesian Lasso could be used within a MICE framework. In particular, the well known Bayesian Lasso proposed by Park and Casella (2008) is a viable option. However, the sparsity parameter introduced by Hans (2010) is what allows for a strictly high-dimensional ($p > n$) data imputation.

MI-CART and MI-RANF Overall, MI tree-based methods performed acceptably in terms of bias, although they rarely excelled. As all other methods, they struggled with large covariance biases. Furthermore, when looking at the focal parameter PRBs in the resampling study, MI-RF was the worst performing MI method, being outperformed even by CC.

In terms of CI coverage, these methods showed mild-to-extreme under-coverage of most parameters in the high-*pm*-high-dimensionality. However, the deterioration in performance was led by the higher proportion of missing cases rather than the increased data dimensionality.

It is also interesting that it made little difference whether the imputation used CART or Random Forests as building blocks, and when the difference was there it was in favour of

the use of the simpler single CART. The use of Random Forests within a MICE algorithm could have been implemented in different ways. In this article, we decided to use Doove et al. (2014) implementations as they are the ones implemented in the popular *mice* package, while other versions are not currently supported by active R-packages. Shah et al. (2014) independently developed another implementation of Random Forests integrated within MICE which was available in the archived R Package *CALIBERrfimpute* (Shah, 2018). The authors are not aware of any empirical evidence showing substantial differences in how the two methods perform.

Single Data Strategies Overall, missForest showed good performances in terms of bias with PRBs smaller than 10% in size for all parameters except item covariances. However, it resulted in severe confidence interval under-coverage of the true parameter values in virtually all of our set ups. Under-coverage coupled with unbiased estimates for univariate parameters means that too little uncertainty is incorporated in the imputation procedure, which is to be expected from a single imputation approach.

Complete Case analysis showed the worst bias performances, with absolute PRBs often bigger than 20%, while occasionally demonstrating good coverage of the true parameter values. This result should be interpreted in light of two considerations. First, coverage close to nominal is a desirable feature of an imputation method only if it accompanied by unbiased parameter estimates, otherwise it is an indication of an imputation algorithm confidently performing poor imputation. Second, Complete Case analysis is by definition forced to used a smaller sample size than all other methods, which inevitably results in inflated standard errors and larger confidence intervals that are more likely to cover the true parameter values, even if their estimate is biased.

6 Conclusions

We investigated a variety of high-dimensional imputation approaches that can deal with large numbers of possible predictors in the imputation models. These methods have the potential to simplify the decisions social scientists have to make when defining which predictors to include in their imputation models. The methods performances in terms of estimation bias and confidence interval coverage of true parameter values were compared with both synthetic and real survey data studies.

We found that *bridge*, a very popular approach to deal with large sets of predictors in the imputation models, is inadequate to deal with strictly high dimensional data set ups ($n < p$). The use of regularized regression within the MICE framework is a powerful tool to automate decisions regarding which variables to include in the imputation models, especially when used exclusively for model trimming (IURR). However, the great performance of these methods comes to a large computational cost that can translates to prohibitive long imputation procedures in real data research.

Finally, the use of PCA to reduce the dimensionality of the data, as a pre-processing step followed by regular low-dimensional MICE imputation strategies, proved to be a fast and effective approach. It was especially effective in preserving relationships between variables with missing values: in simulation study 1 and 2, it returned negligible covariance biases when all other methods failed to. However, its application on real survey data and the need for improvements to imputation accuracy suggest further research is needed.

Limitations and future directions As this work aimed at comparing current implementations of different methods, some limitations to the scope of the simulation and resampling studies were imposed by the current state of development of the different methods. For example, both IURR/DURR and MI-PCA allow imputation of any type of data: IURR and DURR have been developed for categorical data imputation (Deng et al., 2016),

and MI-PCA can be performed with any standard imputation model for categorical data. However, blasso has not been formally developed for multi-categorical imputation target variables yet, which limited the study to working with missing values on variables that are either continuous in nature or usually considered as such in practice. IURR, DURR and MI-PCA could have performed better had they been used in their ordered categorical data implementations, but to maintain a fair comparison ground with blasso, they were implemented with the assumption that the imputed variables are continuous normally distributed.

Furthermore, in real survey data, the missing data mechanism might be non-linear, which would require including interactions between auxiliary variables and polynomial terms in the imputation models. This factor was not taken into consideration as the inclusion of interactions and squared terms in the imputation models has as not been developed to the same extent across the different methods. However, all of the high-dimensional imputation methods considered have great potential to allow the specification of much more complex response mechanisms than traditional methods ones.

Both IURR and DURR could have implemented with different types of penalty formulations. Along with the traditional lasso penalty, Zhao and Long (2016) used elastic net penalty (Zou and Hastie, 2005) and adaptive lasso (Zou, 2006). Although no substantial performance difference between penalties specifications for IURR and DURR emerges from the joint work of Zhao and Long (2016) and Deng et al. (2016), the impact of different types of regularized regression was not explored in the present study.

MI-PCA specification required making a decision on the number of components to extract. In this paper, the authors decided to retain the first components that explain 50% of the total variance in the auxiliary variables, which allowed for substantial dimensionality reduction without relying on too few components. However, this decision is arbitrary and its effect on the imputation accuracy, which was not explored in this study, is an interesting topic for future research.

References

Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37.

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070–1076.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292.

Collins, L. M., Schafer, J. L., and Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Pscyhological Methods*, 6(4):330–351.

D'Ambrosio, A., Aria, M., and Siciliano, R. (2012). Accurate tree-based missing data imputation and data fusion within the statistical learning paradigm. *Journal of Classification*, 29(2):227–258.

de Andrade Silva, J. and Hruschka, E. R. (2009). Eacimpute: an evolutionary algorithm for clustering-based imputation. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 1400–1406. IEEE.

Deng, Y., Chang, C., Ido, M. S., and Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, 6:21689.

Doove, L. L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104.

EVS (2020). European values study 2017: Integrated dataset (evs 2017). GESIS Data Archive, Cologne. ZA7500 Data file Version 3.0.0, https://doi.org/10.4232/1.13511.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hans, C. (2010). Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, 20(2):221–229.

Howard, W. J., Rhemtulla, M., and Little, T. D. (2015). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*, 50(3):285–299.

Immerzeel, T., Coffé, H., and Van der Lippe, T. (2015). Explaining the gender gap in radical right voting: A cross-national investigation in 12 western european countries. *Comparative European Politics*, 13(2):263–286.

Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198.

Köneke, V. (2014). Trust increases euthanasia acceptance: a multilevel analysis using the european values study. *BMC Medical Ethics*, 15(1):86.

Lang, K. M., Little, T. D., and PcAux Development Team (2018). *PcAux: Automatically extract auxiliary features for simple, principled missing data analysis*. R package version 0.0.0.9013.

L'ecuyer, P., Simard, R., Chen, E. J., and Kelton, W. D. (2002). An object-oriented random-number package with many long streams and substreams. *Operations research*, 50(6):1073–1075.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558.

pandas development team, T. (2020). pandas-dev/pandas: Pandas.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21(3):7–30.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, volume 519. John Wiley & Sons, New York, NY.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Savalei, V. and Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(3):477–494.

Shah, A. (2018). *CALIBERrfimpute: Imputation in MICE using Random Forest*. R package version 1.0-1.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American journal of epidemiology*, 179(6):764–774.

Song, J. and Belin, T. R. (2004). Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine*, 23(18):2827–2843.

Stekhoven, D. J. (2013). *missForest: Nonparametric Missing Value Imputation using Random Forest*. R package version 1.4.

Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, FL.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Zhao, Y. and Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*, 25(5):2021–2035.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

7 Appendix

Supplements to simulation study (experiment 2)
    Supplements to resampling study

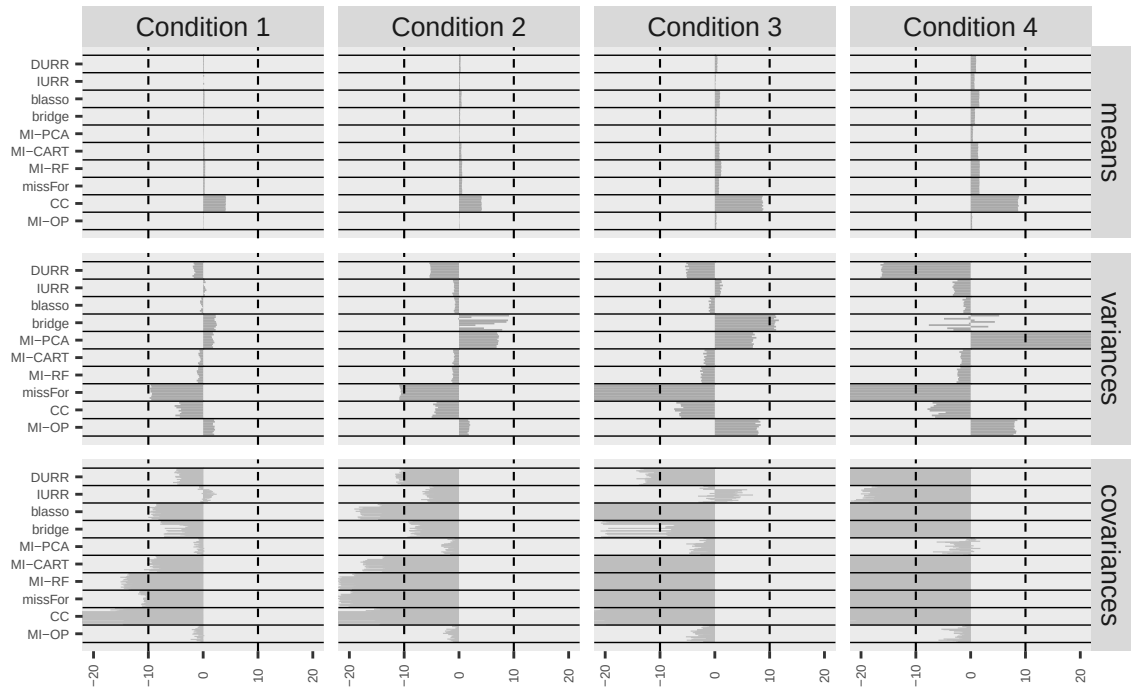

Figure 10: Bias estimation for the means (SB), variances and covariances (PRB) for condition 5 to 8.

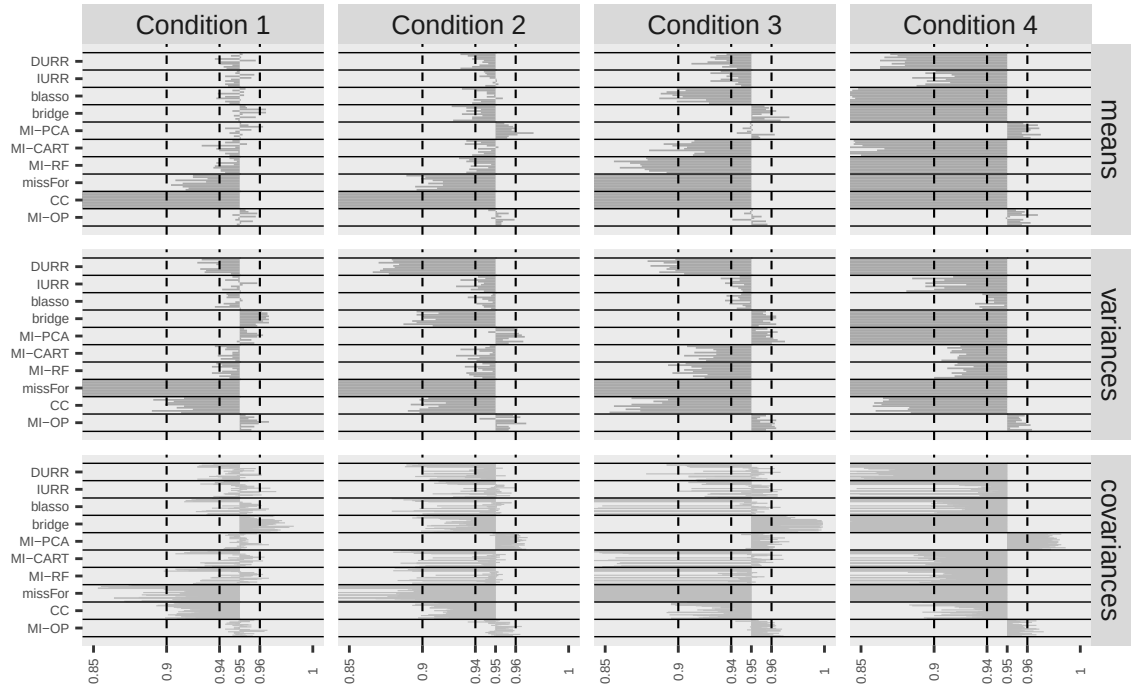

Figure 11: Confidence Interval Coverage (CIC) for the means, variances, and covariances for condition 5 to 8.

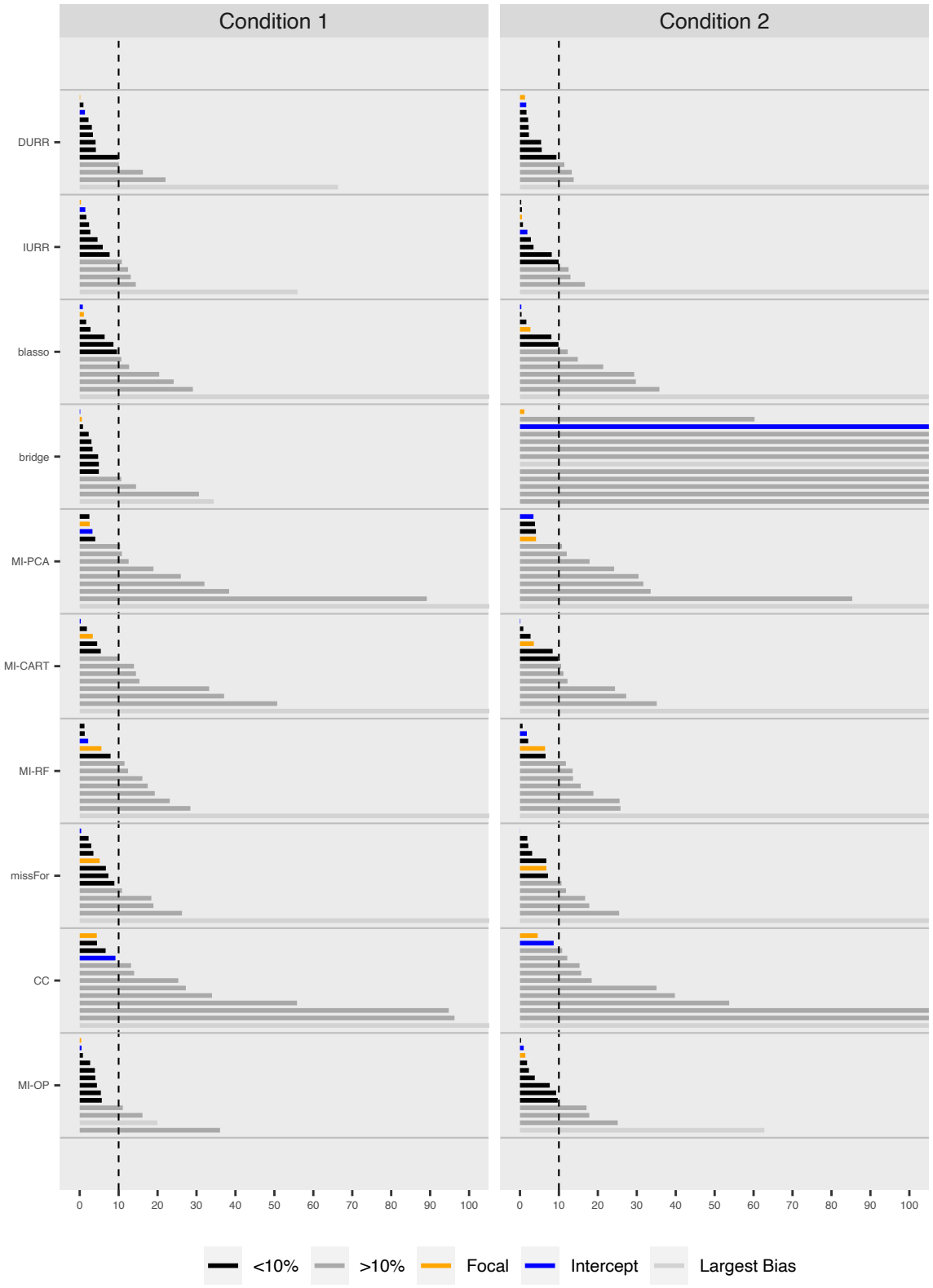

Figure 12: PRBs for all the model parameters in model 2. The order of the bars is based on the absolute value of the PRBs. The values for the intercept, the focal regression coefficient, and the regression coefficient with which most methods struggle (Largest Bias) are highlighted

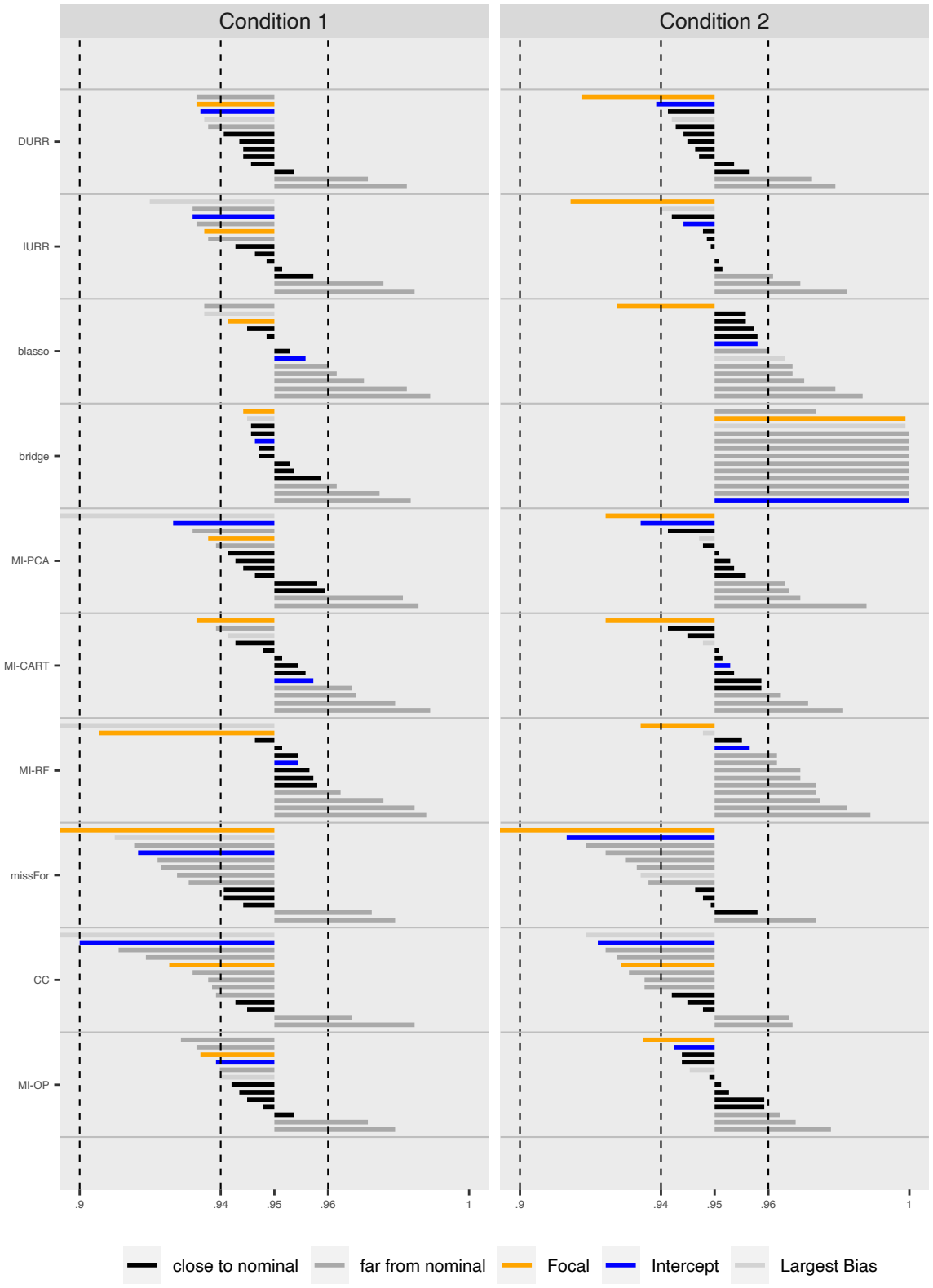

Figure 13: CIC for all model parameter in model 2. The values for the intercept, the focal regression coefficient, and the regression coefficient with which most methods struggle (Largest Bias) are highlighted